Space
and
Naval
Warfare
Systems
Center
San
Diego

BIENNIAL REVIEW 2001

Space
and
Naval
Warfare
Systems
Center
San
Diego

BIENNIAL REVIEW 2001

SPAWAR
Systems Center
San Diego

TD 3117
August 2001

SSC San Diego
53560 Hull Street
San Diego, CA 92152–5001

Approved for public release; distribution is unlimited.

CONTENTS

FROM THE COMMANDING OFFICER AND THE EXECUTIVE DIRECTOR

We are pleased to present the first edition of the Biennial Review, a publication that reflects the innovative research, diverse expertise, and unique capabilities of Space and Naval Warfare Systems Center, San Diego (SSC San Diego).

The Center's vision is to be the Nation's pre-eminent provider of integrated command, control, communications, computers, intelligence, surveillance, and reconnaissance (C^4ISR) solutions for warrior information dominance. This publication features a collection of papers describing significant C^4ISR research and development—a representative sampling of the many technical efforts at the Center in support of our Nation's warfighters. In addition to delivering programs designed to provide C^4ISR capabilities, SSC San Diego pursues a unique range of work in other vital leadership areas, such as ocean engineering, environmental science, marine mammals, and the military application of robotic systems.

SSC San Diego encourages its scientists and engineers to explore new ideas through an independent research program and through the development of advanced technology concepts. This Biennial Review showcases not only the range of scientific and engineering work conducted at the Center but the talent and creativity of our technical staff. We are proud of their efforts.

June 2000 marked the 60th Anniversary of SSC San Diego. We look forward to a bright and productive future, continuing a rich tradition of providing our country's warfighters with C^4ISR technology and systems support—and to contributing to the Navy/Marine Corps goal of achieving information dominance.

We sincerely hope you will find the unique ideas and technical insight presented in this publication both useful and interesting.

R C Kolb

Dr. Robert C. Kolb
Executive Director

Captain Ernest L. Valdes, USN
Commanding Officer

INTRODUCTION

ABOUT SSC SAN DIEGO

Space and Naval Warfare Systems Center, San Diego (SSC San Diego) was established in June 1940 as the Navy Radio and Sound Laboratory, the Navy's first West Coast lab. For more than six decades, the Center, under a variety of names and organizational structures, has provided American warfighters with significant capabilities in the form of weapon systems and electronic technology. Currently, the Center's focus is in the essential area of command, control, communications, computers, intelligence, surveillance, and reconnaissance (C^{4}ISR).

SSC San Diego occupies more than 550 acres at the original site of the Navy Radio and Sound Laboratory—the Pt. Loma peninsula to the south and west of downtown San Diego. Nearby is the Old Town Campus, the latest major addition to the Center's extensive complex of laboratories, test facilities, and offices. To the east, SSC San Diego maintains a small but highly productive branch office in Philadelphia. To the west, strategic locations in Pearl Harbor, Hawaii; Barrigada, Guam; Yokosuka, Japan; and Bahrain provide development, engineering, and fleet support capabilities to Navy units operating throughout the Pacific and Indian oceans.

With a rich tradition of technical experimentation and innovation, strategic locations, unique facilities and network connectivity, a workforce that includes 1,800 scientists and engineers and nearly 800 other technical professionals, and a strong partnership with private industry, SSC San Diego is uniquely qualified to provide the full spectrum of C^{4}ISR capabilities, from basic research and prototype development, to extensive test and evaluation services, through systems engineering and integration, to installation and life-cycle support of fielded systems.

While most of the Center's project work address the requirements of the Navy and the Marine Corps, SSC San Diego also actively supports programs of the Defense Advanced Research Projects Agency, the Army, the Air Force, and the Coast Guard. The nature of the Center's assigned responsibilities necessitates active involvement with a variety of other government agencies at the national, regional, and local levels.

Key to the Defense Department's Joint Vision 2020, the Navy's vision of "Forward...from the Sea," and the Marine Corps doctrine of "Operational Maneuver from the Sea" is information superiority, or, in SSC San Diego's vision, warrior information dominance. SSC San Diego is at the leading edge of technologies that support the transformation of data into information, information into knowledge, and knowledge into understanding. It is clear understanding of the battlespace, and the resultant ability to make and execute effective decisions based on that understanding, that provide the warfighter a decisive advantage over an adversary. As SSC San Diego looks forward into the new century, its overriding challenge will continue to be providing our nation's warfighters the resources they need to achieve battlespace information dominance.

1

Next-Generation Information Systems

■

1 NEXT-GENERATION INFORMATION SYSTEMS

C^4ISR Imperatives—Cornerstones of a Network-Centric Architecture

Clancy Fuzak, William L. Carper, Mary Gmitruk,
James W. Aitkenhead, Tom Mattoon, and
Victor J. Monteleon
SSC San Diego

ABSTRACT

Network-centric operations are military operations that fully exploit the availability of "universal" connectivity. Much discussion of network-centric operations focuses on envisioning future applications of the connectivity. These future applications are a confederation of pieces, not a single unit. The prerequisite for fielding these pieces is an in-place network-centric architecture that can support their implementation. SSC San Diego has identified seven command, control, communications, computers, intelligence, surveillance, and reconnaissance (C^4ISR) imperatives that represent command capabilities needed by military forces. Network-centric architecture requires effectively achieving five of these imperatives. This paper argues the importance of these five, and suggests the value of building technologies to enable these imperatives. This approach allows clearer understanding of the application of technology while assuring consistency with the end objective of network-centric operations.

INTRODUCTION

Network-centric operations have been the focus of serious discussion over the past several years, especially following the wide exposure provided by Admiral Cebrowski's 1998 *U.S. Naval Institute Proceedings* article [1]. Here we take the view that network-centric operations are military operations that fully exploit the availability of "universal" connectivity. Such connectivity can lead to:

- Widespread access to heretofore isolated resources (people, machines, data)
- Improved access to specialized information that has, in the past, been difficult to locate
- Accelerated planning processes
- Introduction of a new dimension to "contact" between opposing forces—cyber contact
- Innovative uses of information
- Development of entirely new ways to work and to think about tasks
- Emergent operational concepts and organizational structures
- Et cetera—think, for example, about emerging Web services and Web uses for personal or business reasons

There will no doubt be many innovative applications for the future network as we build toward network-centric operations. Much discussion of network-centric operations focuses on envisioning these future applications—most of which have not yet been invented. These applications are a confederation of pieces, not a single unit. In fact, that is an intention—the ability to evolve and adapt through "parts upgrade," without having to replace an entire system. The prerequisite for fielding these pieces is an in-place network-centric architecture that can support their implementation. And as is the case with the Web, applications follow infrastructure. Make access simple and widespread, make providing content relatively easy, and someone invents eBay. In this view, "network-centric architecture" provides ubiquitous and universal, timely and "useful" access.

IMPERATIVES FOR C^4ISR

SSC San Diego has identified a set of seven command, control, communications, computers, intelligence, surveillance, and reconnaissance (C^4ISR) imperatives. These imperatives represent command capabilities that have been needed by military forces throughout history and are expected to continue to be needed in the future. While the imperatives are time-independent, the degree to which they can be achieved depends upon available technology.

Dynamic Interoperable Connectivity will provide assured, user-transparent connectivity, on demand, to any desired locations in the "infosphere"—the worldwide grid of people, sensors, military databases, fusion nodes, national resources, and commercial and other non-U.S. information resources.

Universal Information Access will use that connectivity to access strategically located sensors, database servers, and anchor desks. It will provide users, at all levels, with the key information needed to create and share a consistent perception of the operational situation.

Focused Sensing and Data Collection provides the warfighter with the ability to acquire the information needed to allow viewing an area of interest or responsibility at any desired level of fidelity and resolution.

Achieving *Consistent Situation Representation* is the fourth imperative. When all key operational commanders have a consistent situation understanding, tools supporting the fifth imperative, *Distributed Collaboration*, can be used to work effectively together across space and time to plan and execute missions and tasks.

The sixth imperative, *Information Operations–Assurance*, will protect our information and our C^4ISR infrastructure.

Finally, *Resource Planning and Management* provides the mechanisms for effective use of all available resources.

Implementing a network-centric architecture requires effectively achieving several of these imperatives.

NETWORK-CENTRIC ARCHITECTURE

The concept of "ubiquitous and universal, timely and 'useful' access" needs some discussion. The first point we should make is that "access" does not equal information access, which we will discuss later as the imperative for Universal Information Access. In the network-centric architecture, access implies the ability to establish relationships among users. Those relationships must support the users' timeliness requirements. The users might be people, or processes running on machines. Examples of access might be one person phoning another, a person querying a database, a person launching a software process such as an intelligent agent search, a machine process seeking the right human consumer(s) of its information, a sensor establishing relationships with other sensors to triangulate or refine a detection, or a weapon linking to a sensor for guidance purposes.

Some characteristics of the architecture include:

- "Universal" suggests that connectivity must reach everywhere of interest. ("Of interest" is situation dependent.)
- "Ubiquitous" suggests that everything of interest must "plug in" to the connectivity. Plugging in implies some ability to interact with other plugged-in entities under some rules or circumstances—such as appropriate security.
- This "pluggability" implies standards or translators/gateways.

- Where needed access does not exist, it must be "createable" through means such as sensor deployment or establishing connectivity.

Perhaps most importantly, we need to consider "usefulness." We use the term to collectively represent a broad set of attributes that the architecture should support. First, the implementation should be user-centric and intuitive. That is, the implementations should focus on the needs and requirements of users at all operational levels of command, and support those needs in a way that minimizes reliance on specialized skills and training in the use of the architecture elements. The architecture must be adaptable and configurable. These characteristics suggest that the capabilities supported by the architecture will be totally responsive to the user's unique requirements for information to support specific missions, tasks, or functions. Finally, the architecture must be survivable in the face of all types of physical, electronic, or cyber effects, to the same degree that the user and user's physical space are survivable.

With this view, the imperatives Dynamic Interoperable Connectivity, Universal Information Access, and Focused Sensing and Data Collection apply to the architecture directly. The imperative Information Operations–Assurance and the imperative Resource Planning and Management also apply, but in the limited sense of assuring and managing connectivity and access. The Consistent Situation Representation and the Distributed Collaboration imperatives are really customers or applications that utilize the network-centric architecture rather than being fundamental elements of the architecture.

DYNAMIC INTEROPERABLE CONNECTIVITY

Dynamic Interoperable Connectivity is the conduit for all data and information, whether that information moves 15 feet or 15,000 miles. The Dynamic Interoperable Connectivity imperative aims to ensure that the warfighter has reliable and secure access to all needed information. Providing worldwide Universal Information Access requires an integrated global network for gathering and exchanging information. This includes extensive high-capacity landline connections among military users to maintain extensive databases from which warfighters may "pull." It also requires improved in-theater communications for better response to the warfighter's needs, particularly the dynamic movement of imagery and large files.

Not all connectivity users are people. Machines also must exchange data. Connectivity supporting machine data exchange has been accepted Navy practice for the four decades since the introduction of the Naval Tactical Data System and Link-11. Connectivity can involve any number of people and machines, in various locations, as required to accomplish a task. In the future, machines as users must be able to control connectivity on a priority basis.

Dynamic connectivity is flexible, supporting the time-varying needs of users. But it is also economic, supporting the sharing of resources. This allows a given set of resources to serve many times the needs that could be supported by static connections. In addition, individual users generally perform many functions and belong to multiple user communities associated with those functions. The functions may each require only part-time involvement. Connectivity requirements will then track the shifting task involvements.

The future warfighter must have full access to his/her real and virtual area of responsibility, or "operational space." The operational space may be physically small, or global, depending on the user's role. The operational space may be functionally restricted or extend beyond many organizational boundaries (for example, to include allies). Connectivity is required within and among naval nodes,[1] and between both fixed installations and mobile Navy nodes and non-Navy locations worldwide. The non-Navy locations include other Services; other U.S. government installations, facilities, and nodes; Allied forces and locations; commercial and educational entities; and even hostile forces under some circumstances. This diversity is implied by the term *interoperable*. These connectivities require a wide range of attributes. They require varying levels of security, timeliness of connection establishment, timeliness of information transfer, duration requirements for the user–user interaction, robustness against unintentional or intentional disruption, information integrity or accuracy, and simultaneity (conferencing). The varying levels for the many attributes are not set uniquely for a given connectivity—several combinations may be required for any one connection, depending on the circumstances of the moment or on diverse needs of a user performing multiple activities.

Interoperability is critical. When the community of users extends beyond Navy boundaries, interoperability based on the standards of the larger community is required. Supporting interoperability demands the ability to exchange information and commands between users. This, in turn, places demands on all of the underlying procedures, processes, and hardware at every level. Interoperability implies a common (human or machine) language, common security methods and shared "keys," common protocols, and common modulation formats or methods. Where these items are not shared in common, translation mechanisms must be provided.

Now and for the foreseeable future, the number of possible connections and the capacities of those connections between mobile or deployable nodes will fall short of total user demands. Therefore, the command organization will have to allocate available resources to users based on mission and operational needs. Some resources needed to support Dynamic Interoperable Connectivity are inherently limited. Spectrum must be shared among surveillance (both active and passive); navigation; identification, friend or foe; communications; counter-C^3; and weapons systems (soft-kill systems, in-flight missile guidance). Physical space for radios is limited, and today's radio systems (cryptographic device, modem, transmitter/receiver, antenna coupler, antenna) are usually dedicated to a single user or group. A goal for Dynamic Interoperable Connectivity at large nodes (ships, aircraft) is to eliminate dedicated equipment and spectrum. Reducing dedication of equipment and spectrum to single user classes will increase efficiency, expand the number and types of users having communications access at any given time, and reduce costs.

For very small nodes (miniature sensors, hand-held nodes), battery life is critical and energy consumption per bit delivered is a key characteristic. Universal access must be provided in a way that optimizes that characteristic.

[1] The term "node" is used to encompass manned and unmanned locations—including, for example, unmanned aerial vehicles (UAVs) and individual sensors.

UNIVERSAL INFORMATION ACCESS

A revolution in connectivity and distributed computer power is creating a potential for access to information that must be applied judiciously. Universal Information Access describes the interactive processes for information producers and information users (warfighters). The Universal Information Access imperative focuses on the warfighter's need for enough information to act appropriately, but not so much that confusion results. User pull is the "call for as needed" capability that allows the warfighter to access information, only as needed, based on changes in the operational situation. This capability requires robust information servers to support searching by forces deployed anywhere. Repositories of current, pertinent information, located at anchor desks, provide the warfighter with access to seek and receive the right information at the right time. In this paper we focus on information access by the warfighter (person), since machine information access is a subset—relying upon tools (such as intelligent agents) that could also be used by the warfighter.

The Universal Information Access imperative defines ways to meet user information needs for command and control at all levels. Warfighters must be able to access the universe of information without the need for specialized technical skills. The basic capabilities will consist of (1) user pull information transfer, (2) producer push, and (3) preplanned "information ordering."

User pull information transfer is a "call for as needed" capability allowing warfighters dynamic access to information according to mission situations. Warfighters of any rank will access the infosphere.

Producer push distributes information and alerts to customers, allowing command centers to inform and direct warfighters as needed whenever warfighters have insufficient knowledge or indications to formulate a request. Key to producer push is intelligent selection, or screening.

Preplanned information ordering has two components. First, preplanned essential information is assembled by the warfighter (at any command level) before a mission. Preplanned essential information comes from existing databases, which may be fixed in the sense that they are built and maintained independently of any specific mission. Second, information is updated as the mission requires by over-the-air updating.

User interaction is provided through (1) a *warfighter–computer interface*, (2) information assistants, and (3) information control. The warfighter-computer interface is broader in scope than a typical human–computer interface since the warfighter terminal must allow use by an automaton (an information agent) as well as by a human. The great volume of available information demands that warfighters have support in browsing, cataloging, and making sense of information—we call such support *information agents*. Such software assistants will use decision-support algorithms and artificial intelligence to help process the volume and diversity of the infosphere.

FOCUSED SENSING AND DATA COLLECTION

The developing concepts of a revolution in military affairs, or of network-centric warfare, or of operating inside an adversary's decision process, all assume availability of information upon which to base decisions and actions. Tactical decisions must be based on timely understanding, which, in turn, is based upon real-time data extracted from the area of interest.

In this imperative, *sensing* implies gathering data about the physical world through electromagnetic, acoustic/seismic, olfactory, or other measurement means. Sensing might be based on national or strategic systems including satellites and aircraft. It would include platform-based systems fielded on ships, aircraft, or unmanned vehicles. Finally, sensing might be based on deployed or dispersed tactical probes or sensor fields.

The concept of *focused* sensing implies concentration on things of interest, applying available sensing resources to obtain data and information on key subjects and areas. Focusing narrows the scope in one or more of the aspects of location, time, or type, where type refers to the events, features, or elements to be reported.

Data collection implies gathering data about the cyber world, or data about the physical world through means other than direct sensing. This would include extracting from electronic repositories, or manipulations of archived data.

The network-centric architecture extends to the sensor level. Networked sensors can collaborate to refine and enhance their data products. Some sensors will have the ability to act without real-time direction. This may involve refining their focus area, providing selective reports, or even relocating to areas of greater "interest." The primary objective is to provide the data needed by the user, who defines the focus.

INFORMATION OPERATIONS–ASSURANCE

In today's and tomorrow's world of asymmetric threats, protection of our information systems—and the network itself—is essential. Assurance in network-centric environments is less a feature of system operation than it is an empowerment of the users of these systems. Assurance features provide the access controls, authentication mechanisms, confidentiality, and integrity features that enable the users to assert their identity and to access resources in both peer–peer and client–server interactions. Assurance needs to be built into every aspect of a system in a consistent and correlated way. Piecework solutions or post-deployment appendages of assurance features are seldom successful or evolvable. The foundation of security is a clear definition of what is supposed to happen and who is supposed to perform that action. Given a clear definition of what services a system is supposed to offer and who is authorized to avail themselves of these services, assurance can be developed that these services are offered without modification, disclosure, or interruption, and that other unintended actions do not occur.

Assurance features that should be considered in the network-centric architecture include:

- Adaptation to protocol enhancement since reliance on specific protocol features can be short-lived and inflexible;
- Communication routing decisions should offer assurance of correctness. The exchange of routing information is critically important and must be communicated with assurance;
- Assurance features must support the delivery of information to multiple destinations;
- Assurance features must be designed to support joint mission execution and to support interactions with alliances of convenience;
- Interactions should be characterized as peer–peer or client–server, and

be provided. Special considerations must be made to provide services to remotely located users;

- Participants need to be identified in a consistent way throughout a system. A well-structured directory system is essential to coordinate these identifiers;
- Information should flow among people, while control flows should be contained within a site (i.e., the concept of a manager of managers is a bad idea);
- A small number of clearly defined categories of assured services should be supported. All applications that communicate must depend on one or more of these categories of services. Allowing applications to communicate in unique ways makes it very difficult to demonstrate system assurance.

Security services empower the user in the integrated interoperable distributed information sphere of the future—the network-centric architecture. The many aspects of assurance must be carefully crafted into the functional, operational, and structural aspects of information systems to serve future information warfighters.

RESOURCE PLANNING AND MANAGEMENT

Resource Planning and Management provides the tools necessary to identify and allocate resources for any given task or to meet an unplanned contingency. Such tools support effective use of limited resources including personnel, while requiring minimum manpower and skills for their use. Tools are not task-specific, and relate primarily to the planning for and allocation of C^4ISR electromagnetic, information processing, information management, and personnel resources. Resource Planning and Management includes:

- Core services control, including self-diagnostics and healing, data storage and caching, and shared or distributed computing resources;
- The use of modeling and simulation in support of command and control;
- Decision support tools in support of focused logistics, including inventory control models, loss/damage models, and casualty models;
- Sensor tasking and collection management;
- Electromagnetic resources (antennas and other equipment; power levels; signal types and parameters; spectrum) "negotiator"—including communications resource management;
- Information management.

CONCLUSION

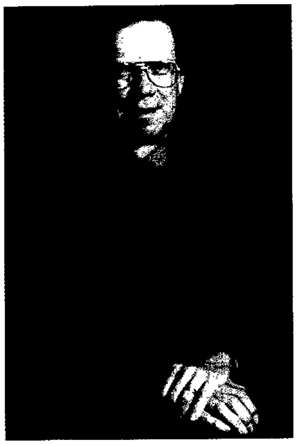
This paper is an attempt to identify the features of an architecture to support evolving and future network-centric operations. Recognizing these required features helps focus our energies on development of the enabling technologies to field the architecture.

AUTHORS

William L. Carper

MS in Electrical Engineering, San Diego State University, 1968

Current Research: System engineering for Naval Space Surveillance System (NSSS) Project.



Clancy Fuzak

Ph.D. in Electrical Engineering, University of Southern California, 1970

Current Research: Concepts and analyses for future naval and joint forces.

Mary Gmitruk

BS in Electrical Engineering, San Diego State University, 1985

Current Research: Roadmapping C^{4}ISR technologies and technology transfer.

James W. Aitkenhead

BS in Physics, San Diego State University, 1973

Current Work: Team leader for the Science and Technology Team; developing new technology for Cooperative Engagement Capability (CEC) process; and participating in the Corporate Initiatives Group (CIG).

Tom Mattoon

BS in Electrical Engineering, University of Idaho, 1970

Current Research: C^{4}ISR architectures and interoperability.

Victor J. Monteleon

MS in Physics and Operations Research, U.S. Naval Postgraduate School, 1966

Current Research: Chair of SSC San Diego's Corporate Initiatives Group (CIG); concepts and architectures for future Navy C^{4}ISR systems; C^{4}ISR vision development.

REFERENCE

1. Cebrowski, ADM A. K., and J. J. Garska, 1998. "Network-Centric Warfare, Its Origin and Future," *U.S. Naval Institute Proceedings*, January, pp. 28–35.

❖

Network-Centric Computing: A New Paradigm for the Military?

LCDR Lawrence J. Brachfeld, USN
SSC San Diego

ABSTRACT

This paper investigates the optimal way to implement ultra thin computer architecture into the existing Information Technology for the 21st Century (IT-21) infrastructure. Factors studied include system architecture, the effect of limited communication capabilities of naval units, changes to current battle group operating doctrine, and the benefits and risks of introducing this new capability to the Fleet.

INTRODUCTION

Network-centric warfare has been defined "as an information-superiority-enabled concept of operations that generates increased combat power by networking sensors, decision makers, and shooters to achieve shared awareness, increased speed of command, higher tempo of operations, greater lethality, increased survivability, and a degree of self synchronization. In essence, network-centric warfare translates information superiority into combat power by effectively linking knowledge entities in the battlespace."[1]

This paper discusses technology known as ultra thin clients (UTCs) and how to make information delivery more reliable and less expensive through the use of "display appliances" using a network-centric computing (NCC) architecture. The NCC approach is targeted to making information delivery simple and inexpensive. It is not a Windows-only or a UNIX-only approach, nor is it a Web browser approach that proposes to replace the inventory of existing legacy commercial off-the-shelf and government off-the-shelf applications with Web applications. The delivery of a wide variety of applications to the user is accomplished by using the network to allow servers to run applications for multiple users. Runtime environment requirements are thus confined to the servers and not propagated to all clients. Clients need only be able to accept redirected screen displays for the applications.

The main points are:

- Servers are categorized as either generic network servers or specific application hosting servers.
- Both categories of servers rely on the concept of being scaleable and taking advantage of technology to service many users.
- Clients are thin or ultra thin, relying on no application-specific code.
- Clients are not dependent on any specific operating system or hardware design.

The NCC deployment is simplified because the applications themselves are not deployed to the clients who represent the greatest number of users. For example, on a carrier with 1000 seats, there is a 1000:1 reduction in application software update costs, one application server vs. 1000 clients. Configuration management is simplified because the UTCs are zero-administration devices; all management is done at the servers. Low total ownership cost naturally follows because of the greatly reduced configuration management and network administration. High service levels are provided to the users because all applications are available over a redundant network architecture with redundant application servers that can be accessed at any UTC on the network by using smart cards.

The NCC architecture shown in Figure 1 depicts how, by using clusters of network and application servers, the display information required can be pushed out to the end user.

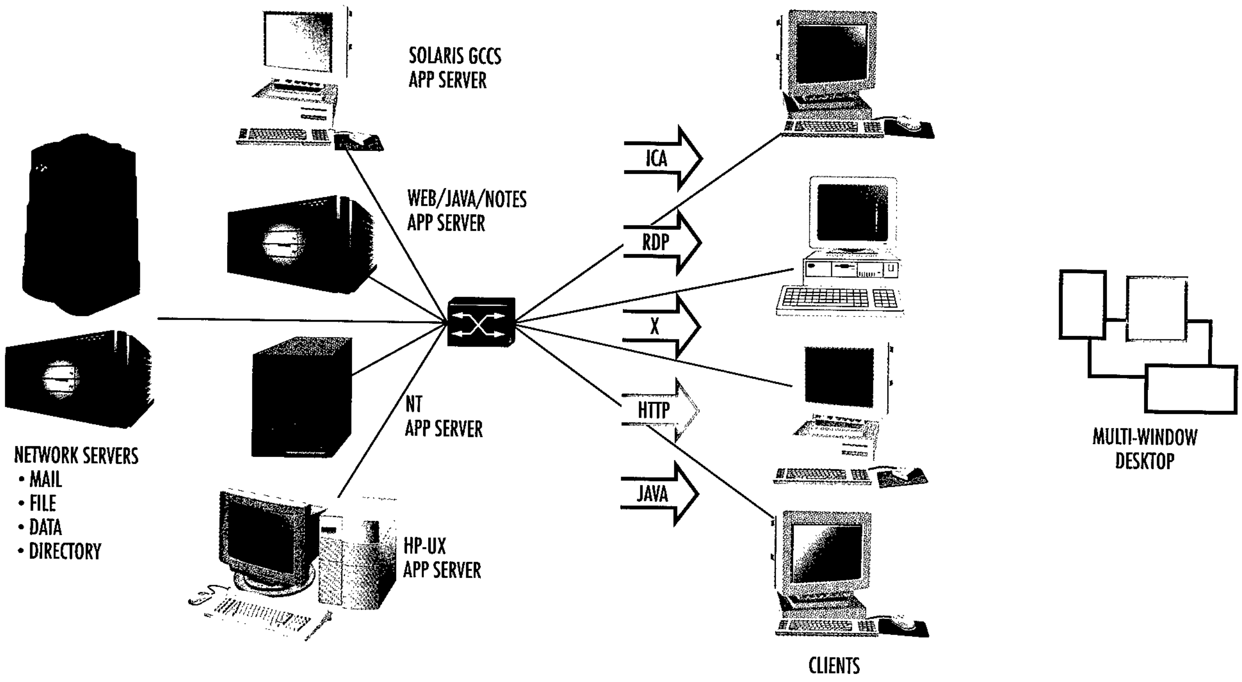


FIGURE 1. NCC architecture.

The NCC architecture has been implemented onboard USS *Coronado's* (AGF 11) Sea Based Battle Lab (SBBL) and has demonstrated the ease and flexibility with which it can be integrated into the existing Information Technology for the 21st Century (IT-21) Integrated Shipboard Network System (ISNS) local-area network (LAN). The current installation consists of 54 UTCs with seamless access to the ISNS backbone for e-mail and office automation, but it also provides access to the Global Command and Control System–Maritime (GCCS–M), GCCS–A (future capability), and Theatre Battle Management Core Systems (TBMCS) (future capability) at the users' desktops, as shown in Figure 2. Figure 3 further depicts the concept of consolidated servers. Additionally, as a natural feature of the UTC, users are no longer "tied" to their PC; they can use their smart card at any of the 54 clients and have full access to all their personal files and network applications.

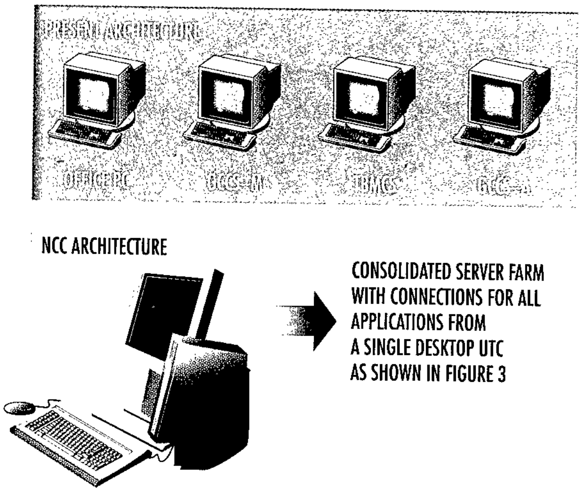


FIGURE 2. NCC desktop environment.

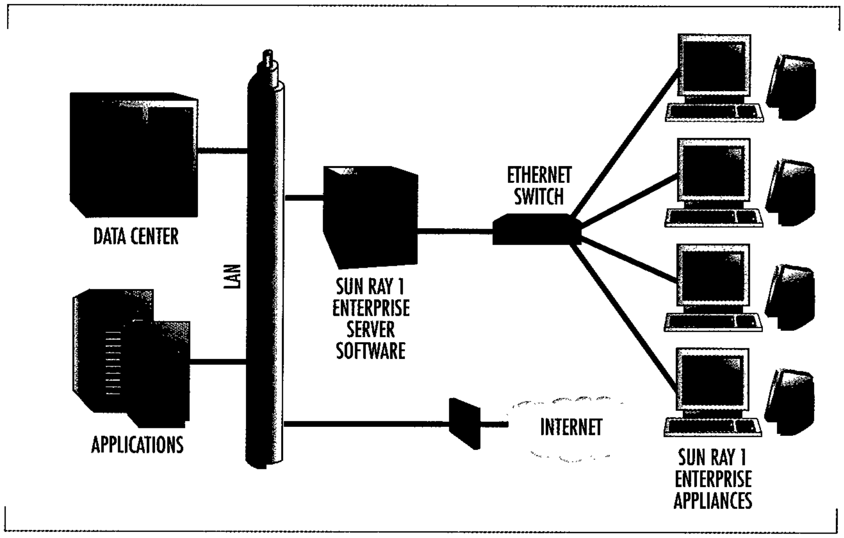

DATA CENTER
APPLICATIONS
LAN
SUN RAY 1 ENTERPRISE SERVER SOFTWARE
ETHERNET SWITCH
INTERNET
SUN RAY 1 ENTERPRISE APPLIANCES

FIGURE 3. Consolidated Server Farm (from Aberdeen Group, September 1999).

As an example, here is a brief illustration of a user's experience with the UTC architecture. Authentication happens once, when a user initially logs on the system by inserting their smart card into the slot on the UTC. There is no wait and no boot-up! The user is immediately connected to the server of their choice and has full access to the programs and files as they were left at the last logoff. The user begins working on a presentation to be given that afternoon and after a few minutes gets a call from the boss asking to see the current presentation. Prior to having a UTC, the user would have had to e-mail the draft plan or save it on a shared network drive; now, without even closing the file, the user removes his or her smart card, walks over to the boss's desk and reinserts the smart card. They are now both immediately looking at the current document, and any changes that are made are saved to the user's file either in a personal directory that only the user can access or on a shared directory to allow for additional collaboration.

As stated in Joint Vision 2020, "the overarching focus of this vision is full spectrum dominance—achieved through the interdependent application of dominant maneuver, precision engagement, focused logistics, and full dimensional protection. Attaining that goal requires the steady infusion of new technology and modernization and replacement of equipment."[2]

To meet this overarching focus with a "steady infusion of new technology and modernization and replacement of equipment" in an environment of shrinking budget resources, a radical shift from the current business model is required. The replacement and modernization of PCs to achieve this vision is impractical; thus, the UTC that never requires upgrading at the end-user location and only requires upgrades at the server level becomes the obvious choice for achieving Joint Vision 2020.

From the installation onboard the *Coronado* SBBL, it has become apparent that the users want more and more applications loaded in this architecture, which allows for the integration of legacy applications that previously required dedicated workstations or PCs. Those applications can now be accessed from any of the 54 UTCs on the network.

This architecture will mark the beginning of a new wave of computing; it is poised to redefine the distributed computing model of the networked fat client PC executing Web-based applications. Although network computing always requires computers, applications, and data, the UTC efficiently repartitions the system and redefines what goes where. By removing all computation and state information from the desktop, we truly have a zero-administration client that can help us achieve Joint Vision 2020 and reduce one of the costliest elements of information technology management.

LCDR Lawrence J. Brachfeld, USN

MS in Information Technology Management, Naval Postgraduate School, 1996

MS in Computer Science, Naval Postgraduate School, 1999

Current Research: Ultra thin client computing; multi-level security, Jini technology integration.

REFERENCES

1. Alberts, D. S., J. J. Garstka, and F. P. Stein. 2000. *Network Centric Warfare*, DoD C^{4}ISR Cooperative Research Program (CCRP), Assistant Secretary of Defense for C^{3}I, (February).
2. Chairman, Joint Chiefs of Staff. 2000. "Joint Vision 2020," (May), http:www.dtic.mil/jv2020

❖

Information Management on Future Navy Ships

Marion G. Ceruti
SSC San Diego

ABSTRACT

This paper discusses issues facing information technology (IT) system developers for Navy ships. Its overall emphasis is on the management of large volumes of tactical information that a ship must collect and process during its missions. Specifically, it describes a very ambitious notion of a future Navy Comprehensive Information Management System (CIMS). Challenges and solutions are suggested for CIMS implementation. Many technical areas of information technology are covered as a set of recommendations for future Navy information systems rather than as an analysis of problems for a particular application. Whereas they reflect the Navy's current and projected needs, many of these recommendations will be possible to achieve only with significant breakthroughs in technology and its applications. Therefore, this paper can serve as a challenge to researchers, engineers, and technology developers in government and industry to find solutions that meet future IT requirements of naval vessels.

INTRODUCTION

Because of its unique capabilities, the U.S. Navy is the primary service to achieve forward-deployed power projection as a means of protecting national interests. In the platform-centric warfare of the past, naval commanders were concerned much more with how to manage the weapons and sensors capabilities onboard their own ships and the information they could acquire than with the total tactical picture of the theater. The main reasons for this were lack of sufficient bandwidth for communications and a lack of technology to fuse, integrate, and display information rapidly. We enter the new millennium with the emphasis on information as an important resource, a condition evident in the current military trend toward network-centric warfare. Enabled by modern database management, networking, and user-interface capabilities, network-centric warfare [1] implies that all platforms in the theater are aware of and contribute to the total information available to all ships, aircraft, and ashore command centers. In some circumstances, a commander could even deploy assets based on another ship.

Network-centric warfare [1 and 2] also implies that the volume of the information available to warfighters on a theater-wide basis will keep growing. This, in turn, also necessitates that engineers provide to Navy commanders a Comprehensive Information Management System (CIMS) that includes the current capabilities of tactical and non-tactical systems. A CIMS also must feature the next-generation technologies that are now the subjects of intensive research and development efforts in the Department of Defense in general and in the Navy in particular. CIMS is not a formal Navy program; rather, it is a generic term to indicate what is expected to evolve over the next decade and beyond.

The challenge facing Navy planners and administrators is how to accomplish this in an atmosphere of cost cutting, limited budgets, and reduced resources. Whereas past military systems relied mainly on specially built equipment that conformed to military specifications, today's Navy and that of tomorrow will feature a greater usage of commercial off-the-shelf hardware and software. This trend will enable not only cost savings but also the use of new products and services of industry to maintain the leading edge in technology for the warfighter. (See, for example, [3]).

Also consistent with the policy of cost savings, next-generation Navy ships will have fewer personnel. Sailors will need to learn multiple jobs and become familiar with multiple tasks, in addition to what they are doing today. This implies that more automation will be necessary in all areas, especially automation in training systems, such as Web-based training. More than ever, tomorrow's Navy will learn how to accomplish more with fewer resources.

CIMS CAPABILITIES

The ideal CIMS will provide the following capabilities and address the following topics:

- Database integration and knowledge-base integration
- Knowledge-base integration with databases
- Database and knowledge-base standards and refresh for these standards
- Maintenance of security during database integration
- Data standardization to facilitate database and knowledge-base integration
- Data warehouse technology and data warehouse software refresh
- Data preprocessing and cleansing prior to storage in data warehouse
- Data mining that includes mission-directed Web searches
- Data-mining tool refresh
- Enhanced data-fusion technology
- Advanced data storage systems
- User-friendly database and knowledge-base access
- Database and knowledge-base management, including correct database management system (DBMS) and knowledge-base management system selection and refresh of commercial off-the- shelf and government off-the-shelf and software
- Regular updates of standard command and control systems, such as the Global Command and Control System–Maritime (GCCS–M) [4] and the databases that support them
- Periodic assessment of data storage requirements and plan to meet future needs
- Use of intelligent agents in conjunction with data warehouse, databases, and knowledge bases
- Knowledge- and data-replication to avoid a single point of failure
- Subsystem to provide situational awareness
- Computer network information on all offensive efforts
- Information-service "reach-back" to networked ashore capability
- Information warfare activity integration
- Integration of intelligence and security information

CHALLENGES AND SOLUTIONS FOR CIMS IMPLEMENTATION

The Navy must overcome many obstacles before the completion of a CIMS. This section describes some of these obstacles and challenges [4] that Navy information systems engineers will encounter.

Data Fusion

Naval forces need to link and fuse in real time more sensor data from a wide variety of sources. This implies a requirement for a modular, open-systems environment in which various data fusion engines can be inserted or deleted. Meeting this requirement necessitates an unprecedented data fusion effort for sensors on aircraft, unmanned airborne vehicles, satellites, and precision weapons of all U.S. and allied forces. The Navy will fuse information, or will used finished fused data products, from other services and allies in the common operating picture. The CIMS ideally will accommodate any sensor input—a situation that is very open-ended. Therefore, one challenge is for the U.S. Navy to know when this requirement is satisfied, especially when the Navy has no direct control over the interface designs of sensors from the other services and allies. (For more information on the joint vision, see [5 and 6]).

Distributed Database Components

Data will be collected and integrated. For example, the CIMS will contain the biological and chemical sensor information that will be integrated. Engineers will need to develop metadata [7] documentation of database systems components with an explanation of the relationship between components (e.g., how their data elements are subsets or a superset, etc., of the integrated databases) that support major existing systems, such as GCCS-M. Database access efficiency depends on the hardware, the DBMS, the operating system, and the relative priorities of competing tasks. Thus, the CIMS will feature a modernized version of a distributed, federated database. (See, for example, [1, 8, 9, and 10]).

Optimized Data Structure

The establishment of an information warehouse in a data management system for all users is not enough to guarantee an optimized data structure. Therefore, engineers must consider all of the factors necessary to achieve an optimized data structure. Also, engineers must provide to the users (e.g., sailors) an online document that will explain the operations for which the data structure will be optimized. For example, a data structure optimized for retrieval performance will not be optimized for data storage performance and vice versa [11]. The documentation will list the advantages and disadvantages of the particular data structure selected for implementation. This information is generally not present in current command and control database systems in any comprehensive sense.

Data and Database System Standardization

The CIMS will feature data standardization that is needed, not only for sensor-data fusion, but also for other aspects of data integration. The CIMS will contain an up-to-date reference list of all necessary and germane data standardization documents. The Defense Information Systems Agency (DISA) has instituted the Defense Information Infrastructure Common Operating Environment (DII COE) as an essential element for inter-service interoperability [4]. The DII COE includes the Shared Data Environment. The CIMS will comply with DISA's standards at each level of DII COE.

Data Aggregation

The CIMS will provide access to distributed legacy databases through a user interface, which is a step toward data aggregation. However, this is insufficient to guarantee uniform data services to all active components. It is only a step on the way to data integration and not data integration in its entirety. The challenge that the Navy faces here is to determine all steps in the information integration process, including data aggregation and addressing any security implications that this aggregation creates [4] on a resources-available basis.

Information Integration

Information Integration Analysis

Extensive analysis is necessary to integrate and present clear and non-redundant information. The Navy will face the challenge of ensuring that the CIMS will be based on the analyses that have been performed, considering the cost and security implications. The ideal CIMS will use what engineers have learned from others' experiences in information systems integration so it can present clear, useful, timely, and non-redundant information to its users.

Information System Integration Details

The Navy will need to describe and document its integration approach, including how much integration can be completed given the financial constraints. To accomplish a successful CIMS, engineers will need to provide details of how information systems integration will be performed on all levels, including semantic and data levels of integration. The engineers will become familiar with the integration methodology and the algorithms used to accomplish it. A list of integration priorities must be developed because all desired integration tasks cannot be performed in a reasonable timeframe and within budget [12].

Online documentation will describe the database integration strategy and or methodology with enough detail so that personnel who are not computer experts will know that the integration method will result in the required seamless database interfaces and will include integration on all levels. Data residing at different decision centers will not be consistent automatically. Therefore, the CIMS will need to be able to identify and resolve the inconsistencies. (See, for example, [9 and 13]). The integration method and architecture will be specified. The level of integration in the CIMS will be specified so that the user will know what the developers could accomplish at the allocated funding level. Ideally, the CIMS should be integrated on three levels: platform, syntactic (data model), and semantic [9 and 13].

Integration Methods and Large-screen Displays

Large-screen displays are a common feature of modern command centers. Large-screen displays can facilitate error detection on an *ad hoc* basis, but they cannot substitute for a thorough database integration effort. To reduce inconsistencies in the data, more automated methods are needed and specific algorithms should be utilized. The CIMS should provide a description of all integration methods that will be used before giving users a possible means (but not a systematic method) to notice data inconsistencies via large-screen displays.

Data Cleansing

The ability of an information warehouse, a common backbone, and a large-screen display to increase reliability and consistency is only as good as the integration and data cleansing [14] that has taken place in the data sources. This integration and data cleansing must be performed before taking the following steps:

- Installing the data in the warehouse
- Making data available on a common backbone
- Displaying them on large-screen displays

Ideally, only clean and consistent data will be stored in the information warehouse. However, few if any databases of appreciable size have ever had totally clean and consistent data.

Information Warehouse

In the ideal CIMS, an information warehouse provides a complete source of warfighting information and knowledge to all echelons. To accomplish this, engineers will have to define metrics to evaluate the completeness of warfighting information and knowledge in the information warehouse. They will need to test and evaluate the ability of the information warehouse [15] to deliver information efficiently to the user at each stage of compliance. For example, it may be possible to provide a 70% solution at time, t, and an 80% solution at time, t+x. The CIMS will function best if database administrators load all warfighting information and knowledge into the information warehouse in well-defined stages. A difficult challenge to engineers will be to determine how all data systems will be integrated into a single information warehouse. A common metadata repository must be part of the data warehouse to support the CIMS and the common operating picture.

Knowledge Management

Knowledge Standards and Knowledge Management

Commercial, open-system standards will contribute to an affordable and information system architecture designed to accept upgrades efficiently. Database management services with Relational Database Management Systems and with Object Relational Database Management Systems, such as Open Database Connectivity, are well known. However, standards as they apply to existing capabilities and equipment are insufficient. Because knowledge centricity is an important feature of future ships, the information processing standards will need to include the emerging knowledge standards, such as Open Knowledge-Base Connectivity (OKBC) [16], Knowledge Interchange Format, or Knowledge Query Markup Language [17]. OKBC is the knowledge analog of Open Database Connectivity. Standards need to support open data and information exchange architecture. The CIMS will support this criterion by including the class of standards to address knowledge interchange. The current knowledge standards will evolve to higher levels during the coming decade. Therefore, the CIMS will be evaluated for a periodic refresh of knowledge standards as new ones emerge. These standards will contribute to database and knowledge-base integration, including the integration of ontologies necessary to support future artificial-intelligence technology in the knowledge management system(s) of the CIMS.

Common Ontology

The CIMS will have a common ontology and a knowledge base derived from it that will be accessible to all users over the network. This ontology will be necessary to enable the semantic integration that knowledge centricity implies [17]. In addition to OKBC, a common ontology and the tools to merge ontologies and knowledge bases (of other services and allies) are necessary pieces of the puzzle [16 and 17]. The ideal CIMS will include the complete integration of knowledge bases and databases into a seamless common operating picture. The Defense Advanced Research Projects Agency has sponsored the High-Performance Knowledge Base program, which produced results that can contribute directly to information-systems and ontology integration problems. (See, for example, [16, 17, 18, and 19]). The CIMS also will make a common ontology available to intelligent software agents. The Navy's challenge in this area is to identify the correct ontologies for integration and to include all relevant concepts in the unified ontology.

Data Mining

Data fusion processing and planning processing are necessary but insufficient by themselves to ensure functional knowledge-centric decision centers. Tactical data mining will be a capability exploited on future Navy ships. The CIMS will assist users to perform the steps of data mining to be carried out on each ship. The CIMS also will assist users in determining the desired outcomes of data mining for a particular task and the tactical data-mining tools required to complete the task. The CIMS will integrate the outputs of the intelligent software agents and coordinate the behavior of the intelligent software agents with each other with the output of the tactical data-mining tools to augment the knowledge base. Promising current approaches to data-mining problems [20, 21, 22, 23, 24, 25, and 26] in the area of command and control [20, 21, 24, and 25] range from Bayesian networks for data-classification tasks [21] to knowledge mining with randomization and features to overcome the knowledge-acquisition bottleneck [25].

Mission-directed Data Mining

Although Internet connectivity is common in today's Navy and will be part of the total communications package, a specific need for this type of connectivity has been identified to support cryptologic and information-operations, mission-directed Web searches. The CIMS will enable sailors to implement cryptologic and information-operations, mission-directed Web searches, and to integrate the information obtained from such searches with other data sets already in the database where appropriate [1].

Data-mining Technology Upgrades

Tactical data mining is not a reality today. The whole data-mining process as we know it typically takes too long to be accomplished in seconds and is therefore not yet suitable for tactical, real-time applications. However, in the coming decade, tactical data mining may be not only possible, but tools to accomplish it may be modular, commercial off-the-shelf, user-friendly, and compliant with Department of Defense standards. Therefore, the CIMS will include technology refresh in the area of data mining to enable this new and developing technology to contribute significantly to knowledge centricity.

Information Operations, Efficiency, and Security

Information Operations

The information management and information integration activities will be coordinated with the information operations activities to provide efficient and seamless information services. An ideal CIMS will be able to handle smooth interoperation and conflict resolution between these activities.

Situational Awareness

Situational awareness relates to the common operating picture, the common tactical picture, etc., that will be available to all Navy personnel in the theater. Tactical decision-makers on future ships will have an adequate situational awareness about their operational posture (friendly, hostile, and neutral) in the electromagnetic spectrum, in the computer network environment, and in other domains such as the psychological, cultural, and environmental "pictures." Inherent in this requirement is the need for appropriate decision aids, algorithms, displays, simulation tools, etc., to provide situational awareness in the information-operations arena.

Computer Network Exploit and Attack

A Strike Force Commander must be aware of all offensive efforts that may affect the strike (hard kill or soft kill). The future CIMS will need a specific capability to provide the Strike Force Commander (and other Strike Force Commanders in the theater) information on all offensive efforts to avoid overkill of targets that could cause the unnecessary expenditure of scarce and/or limited ordnance resources.

Reach-back

To reduce the most expensive cost factor (payroll), personnel limits have been specified for future ships, with the expectation that functions can be moved ashore and future ship operators can "reach back" for what they need. To make sure that future ship personnel will have all the information services they need at the same level of reliability, these supporting shore services will need to be more secure, robust, redundant, and capable than they are today. An ideal CIMS will need to meet all of the information system requirements, either onboard the ship, in the theater, or on shore.

Information Warfare Activity Integration

The Navy divides information warfare into two categories: (1) offensive (information attack) and (2) defensive efforts (information protection and assurance). The ideal CIMS will integrate these functions aboard future ships, for both traditional information systems, e.g., tactical communications, message traffic, voice, etc., and those associated with the computer network environment.

Levels of Security

Secure database technology is now available. The CIMS will feature multi-level security (MLS), which will address issues such as MLS vs. network security, network security vs. secure operating system and/or secure DBMS, etc. Security needs to be implemented at all levels to preclude a weak link in the security chain. Network security is not enough. Most of security is enforced on the network in a network-centric security system. The CIMS will provide security at the operating systems and database management systems level.

SUMMARY

The information management systems on future Navy ships will provide rapid access to fused and integrated data and knowledge to meet the ever growing needs of tomorrow's warfighter at sea. The technology now in the research and development stages will make a valuable contribution to enhance the capabilities of our naval and joint forces throughout the coming decades.

ACKNOWLEDGMENT

The author thanks the Naval Sea Systems Command for financial support of this effort.

Note: This paper will also appear in the *Proceedings of the AFCEA 18th Annual Federal Database Colloquium and Exposition* on "Database Scalability, Quality, and Interoperability," 28–30 August, 2001, San Diego, CA.

Marion G. Ceruti

Ph.D. in Chemistry, University of California at Los Angeles, 1979

Current Research: Information systems analysis, including database and knowledge-base systems, artificial intelligence, data mining, cognitive reasoning, software scheduling and real-time systems; chemistry; acoustics.

REFERENCES

1. Ceruti, M. G. 1999. "Web-to-Information-Base Access Solutions," *Handbook of Local Area Networks*, chapter 4-5, pp. 485–499, Auerbach Publications, Boca Raton, FL.
2. Ceruti, M. G. 2001. "Mobile Agents in Network-Centric Warfare," *Proceedings of the 5th IEEE International Symposium on Autonomous Decentralized Systems* (IEEE ISADS 2001).
3. Johnson, J. L., CNO, U.S. Navy. 1997. "Forward . . . from the Sea: The Navy Operational Concept," http://www.chinfo.navy.mil/navpalib/policy/fromsea/ffseanoc.html
4. Ceruti, M. G. 1998. "Challenges in Data Management for the United States Department of Defense (DoD) Command, Control, Communications, Computers and Intelligence (C^4I) Systems," *Proceedings of the 22nd Annual International Computer Software and Applications Conference, IEEE COMPSAC'98*, pp. 622–629. Also, see http://www.disa.mil/disahomejs.html
5. Defense Technical Information Center, Joint Vision 2010, http://www.dtic.mil/jv2010/
6. Defense Technical Information Center, Joint Vision 2020, http://www.dtic.mil/jv2020/jvpub2.htm
7. Foss, R. and R. Haleen. 1997. "An Environment for Metadata Engineering," *Proceedings of the 14th AFCEA DoD Database Colloquium '97*, pp. 75–91.
8. Ceruti, M. G. and S. A. Gessay. 1998. "White Paper on the Next-Generation Data-Access Architecture for Naval C^4I Systems," *Proceedings of the 15th AFCEA Federal Database Colloquium '98*, pp. 451–472.
9. Ceruti, M. G. and M. N. Kamel. 1994. "Semantic Heterogeneity in Database and Data Dictionary Integration for Command and Control Systems," *Proceedings of the 11th AFCEA DoD Database Colloquium '94*, pp. 65–89.
10. Putman, J., B. M. Thuraisingham, W. Chitwood, and M. G. Ceruti. 2000. "Experience in Developing an Information-Sharing Environment in a Large Government Enterprise Using WWW, Federation, Business Components, and Data Warehousing Technologies," *Proceedings of the 17th AFCEA Federal Database Colloquium and Exposition*, pp. 229–244.
11. Ceruti, M. G. 1996. "Development Options for the Joint Maritime Command Information System (JMCIS) Specialized Data Servers," *Proceedings of the 13th AFCEA DoD Database Colloquium '96*, pp. 217–227.
12. Ceruti, M. G., B. M. Thuraisingham, and M. N. Kamel. 2000. "Restricting Search Domains to Refine Data Analysis in Semantic-Conflict Identification," *Proceedings of the 17th AFCEA Federal Database Colloquium and Exposition*, pp. 211–218.
13. Ceruti, M. G. and M. N. Kamel. 1998. "Heuristics-based Algorithm for Identifying and Resolving Semantic Heterogeneity in Command and Control Federated Databases," *Proceedings of the IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'98)*, pp. 17–26.
14. Ceruti, M. G. and M. N. Kamel. 1999. "Preprocessing and Integration of Data from Multiple Sources for Knowledge Discovery," *International Journal on Artificial Intelligence Tools (IJAIT)*, vol. 8, no. 2 (June), pp. 152–177.
15. Malloy, K. 1998. "Data Warehouse Requirements: Scalability, Availability and Manageability," *Proceedings of the 15th AFCEA Federal Database Colloquium '98*, pp. 569–577.
16. Chaudhri, V. K., A. Farquhar, R. Fikes, P. D. Park, and J. P. Rice. 1998. "OKBC: A Programmatic Foundation for Knowledge Base Interoperability," *Proceedings of the 15th National Conference on Artificial Intelligence*. (Also as KSL Technical Report KSL-98-08).

17. Ceruti, M. G. 1997 "Application of Knowledge-base Technology for Problem Solving in Information-Systems Integration," *Proceedings of the 14th AFCEA DoD Database Colloquium '97*, pp. 215–234.
18. Fikes, R., A. Farquhar, and J. P. Rice. 1997. "Tools for Assembling Modular Ontologies in Ontolingua," *Proceedings of the Fourteenth National Conference on Artificial Intelligence.* (Also as KSL Technical Report KSL-97-03).
19. Lin, A. D. and B. H. Starr. 1998. "HIKE (HPKB Integrated Knowledge Environment)–An Integrated Knowledge Environment for HPKB (High Performance Knowledge Bases)," *Proceedings of the IEEE Knowledge and Data Engineering Exchange Workshop, KDEX'98*, pp. 35–41.
20. Ceruti, M. G. 2000. "The Relationship Between Artificial Intelligence and Data Mining: Application to Future Military Information Systems," *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, p. 1875.
21. Ceruti, M. G. and S. J. McCarthy. 2000. "Establishing a Data-Mining Environment for Wartime Event Prediction with an Object-Oriented Command and Control Database," *Proceedings of the Third IEEE International Symposium on Object-oriented Real-time Distributed Computing, ISORC2K*, pp. 174–179.
22. Ganti, V., J. Gehrke, and R. Ramakrishnan. 1999. "Mining Very Large Databases," *IEEE Computer*, vol. 32, no. 8 (August), pp. 38–45.
23. Han, J., L. V. S. Lakshmanan, and R. T. Ng. 1999. "Constraint-Based Multidimensional Data Mining," *IEEE Computer*, vol. 32, no. 8 (August), pp. 46–50.
24. Ramakrishnan, N. and A. Y. Grama. 1999. "Data Mining: From Serendipity to Science," *IEEE Computer*, vol. 32, no. 8 (August), pp. 34–37.
25. Rubin, S. H., M. G. Ceruti, and R. J. Rush, Jr. 2000. "Knowledge Mining for Decision Support in Command and Control Systems," *Proceedings of the 17th AFCEA Federal Database Colloquium and Exposition*, pp. 127–133.
26. Thuraisingham, B. M. and M. G. Ceruti. 2000. "Understanding Data Mining and Applying it to Command, Control, Communications and Intelligence Environments," *Proceedings of the 24th IEEE Computer Society International Computer Software and Applications Conference, COMPSAC 2000*, pp. 171–175.

❖

Object Model-Driven Code Generation for the Enterprise

William J. Ray
SSC San Diego

Andy Farrar
Science Applications International Corporation (SAIC)

ABSTRACT

This paper discusses the benefits of using a code generator to synthesize distributed, object-oriented servers for the enterprise from object models. The primary benefit of any code generator is to reduce the amount of repetitive code that must be produced, thus saving time in the development cycle. Another benefit to our approach is the ability to extend the services generated, enabling the code generator to act as a force multiplier for advanced programmers. Having a code generator synthesize complex code dealing with concurrency, replication, security, availability, persistence, and other services for each object server will ensure that all servers follow the same enterprise rules. Also, by using a code generator, developers can experiment more easily with different architectures. One of the final benefits discussed in this paper is that when using a code generator for the data layer of enterprise architecture, changes in software and evolving technology can be handled more readily.

INTRODUCTION

Joint Task Force–Advanced Technical Demonstration (JTF–ATD) was a Defense Advanced Research Projects Agency (DARPA) project in the field of distributed, collaborative computing. In a typical JTF command hierarchy, the critical people, relevant data, and their supporting computers are geographically distributed across a wide-area network. This causes many problems that would not exist if they were all in the same location. The goal of JTF–ATD was to make it easier for people to work together. A system that facilitated the sharing of data and ideas without compromising security, timeliness, flexibility, availability, or other desirable qualities was needed. After experimentation with numerous architectures and implementations, the JTF–ATD concluded that an enterprise solution to data dissemination and access was needed. It also became apparent that the different types of data needed to support JTF missions were as ubiquitous as the missions themselves. Therefore, planning systems would need the ability to associate previously unknown data elements to their plan composition. A distributed, object-oriented design held the most promise to meet these goals.

Unfortunately, building distributed, object-oriented data servers with the complex infrastructure to support enterprise solutions was costly and time consuming. JTF–ATD built the Next Generation Information Infrastructure (NGII) toolkit to address this problem. The NGII toolkit allows developers to code generate object-oriented data servers in days rather than months. The NGII code generator synthesized complex code dealing with concurrency, replication, security, availability, and persistence for each server, thus ensuring that all servers followed the same enterprise rules. The NGII toolkit and its descendant, Quava, are widely used by many projects today to help generate distributed, object-oriented servers with the intelligence to act in concert across the enterprise. Quava is available to the public and can be downloaded at http://www.saic.com/quava/.

RELATED WORK

Work related to the topics discussed in this paper includes research in program synthesis, code generation, software prototyping, software reuse, software engineering, and software maintenance.

Although much of the research in the fields of program synthesis and code generation deals mainly with optimization, the process of generating code for optimizing digital signal processors (DSPs) or machine language has many similarities to the generation of code for an enterprise data layer. In earlier work, several researchers have generated code from descriptive languages or object models [1, 2, 3, and 4]. Whether the code generated was machine language or code that needed to be compiled is not material to the process of generating the code from a more abstract foundation.

Some researchers even took the generation of code a step further to aid in the creation of control code for multiple processes. In the Computer-Aided Prototyping System (CAPS), code is generated from a more abstract language to simulate a real-time system [5 and 6]. Attie and Emerson synthesized concurrent programs from temporal logic specifications [7].

Software reuse has always fallen short of its lofty goals. The reasons cited for its failure are too numerous to list [8]. Some of the most promising work to help reach the goals of software reuse involves a hybrid approach of program synthesis by making use of reusable code components and code generation [9]. This approach is the one taken by the tools described in this paper.

CODE GENERATOR

Quava provides application developers with an Integrated Generation Environment (IGE) that allows them to convert engineering designs from Computer-Aided Software Engineering (CASE) tools (e.g., Rational Rose, Oracle Designer, etc.) into Unified Modeling Language (UML) encoded design objects. Quava can then generate implementation code that can incorporate Common Object Request Broker Architecture (CORBA), Remote Method Invocation (RMI), Component Object Model (COM), or Java 2 Enterprise Edition (J2EE) services. The developer has complete control over which services, architecture, and language to use for their application.

Design

The Quava system is composed of four basic pieces (Figure 1).

The first piece, the repository adapter, imports data and can communicate with commercial off-the-shelf (COTS) modeling tools, such as Rational Rose or Designer 2000, or read models stored in the Object Management Group's (OMG's) XML Metadata Interchange (XMI) file format. XMI is key to interoperability with other COTS modeling tools. The repository adapter imports a model, which is then instantiated as a UML 1.3 metamodel. Internally, Quava can store its UML

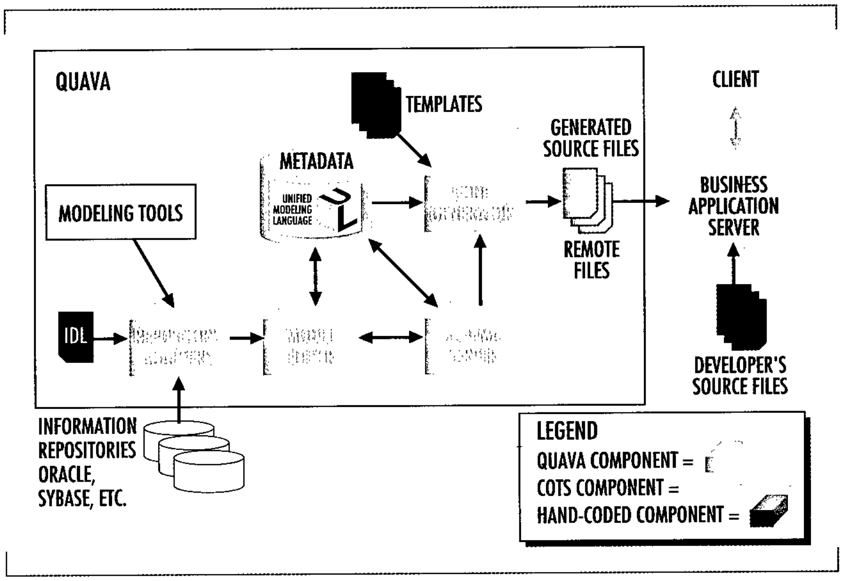


FIGURE 1. Code generation system.

objects either in an XMI flat file or to a UML server, called the Schema Server, for enterprise-wide sharing of models.

The second piece is a tool for altering the UML model. While Quava is not a modeling environment, we did allow for model editing because many COTS tools only support older versions of the UML standard, and many do not support the kinds of additional modeling information designers may want to express. Quava provides the Model Editor, which allows a user to go in and change or add information to the UML model. One example of this is the mapping of one model to another. This is very common when mapping from an application object model to a database model. Model-to-model mapping can also occur between different UML models to automatically generate interface code from a specific model to a shared model. In creating any additional modeling information, Quava still maintains the UML standard by only using UML metamodel objects to represent the additional information. This allows changed models to remain compatible with other COTS modeling tools.

The third piece is a set or sets of templates that guide and direct the generator to precisely what code to produce. Quava differs from many code generators that are used to produce code for a specific COTS tool or environment. Quava users can change or add new templates to allow production of any type of output in any language. The templates are written in either ECMAScript, which is a standardized version of JavaScript, or in Java. The templates allow for maximum flexibility and provide a mechanism for the users to define both the code output and the process flow the generator takes through the model.

The fourth piece is the generation engine. The generation engine pulls in a UML model and then proceeds to apply the selected set of templates against the different elements of the UML model. Processing continues until all selected templates have been processed against the model. Finally, unlike many other tools, the code produced is not tied to Quava in any way and can be imported into whatever development environment the user typically uses.

TEMPLATES

The templates that drive code generation are the key to both the generation's output and the level of control the user has over the generation process. During the course of experimentation with the model-driven code generation approach, we focused on three main issues. The first issue was identifying which types of services lend themselves to model-based code generation. The second issue was how code generation could help with the composition of services in a large-scale architecture, and the third issue was how easily the templates could be extended or new templates added.

Types of Services

To identify which services best lend support to a model-based code generation approach, we focused on where developers spend most of their time. Current software products allow users to generate skeleton code for different architectures, but this code is limited to just a single architecture and does not help with any of the actual logic of the objects. So, where would users get the most "bang for the buck"? Architectural services. Architectural services came to the forefront because they require the developer to implement additional functionality into each object in the schema in support of the service. For example, an Extensible Markup Language (XML) streaming service may provide a class library for creating the stream and sending and receiving a stream, but the objects within the system will need to implement a method to serialize their attributes to an XML stream. This type of service, where knowledge of the model can reduce the amount of work a developer has to do, is exactly where the code generation process fits in. Below is a very simple ECMAScript template for generating a method to serialize an object to an XML stream.

```
/************************************************************/
// Xml Example/
function writePackage(modelhdl)
{
 var i, interfaceName;
 // Get All the element in this model
 classesList = modelhdl.getOwnedElement();
 // Loop through each element in the model
 for(i = 0; i < classesList.size(); i++) {
   // GLOBAL class object
   xmlClassObj = classesList.elementAt(i);
   // If it's a class then process it otherwise look for nested packages
if(xmlClassObj.getClass().getName() =="mil.darpa.ngii.uml.umlClass")
     writeClass(xmlClassObj);
     else                              if(xmlClassObj.getClass().getName()==
"mil.darpa.ngii.uml.Package")
     writePackage(xmlClassObj);
   }

   }

  /*************************************************************
*****************/
  /**
   * Write the class structure: header, attributes, and footer.
   */
  function writeClass(xmlClassObj)
  {

   myXMLFile.writeln("public void writeToXML(StringWriter out)");
   myXMLFile.writeln("{");
   myXMLFile.writeln(" out.write(/"<class>/");");
   myXMLFile.writeln("
out.write(/"<classname>"+xmlClassObj.getName()"+</classname>\n/");
");
  myXMLFile.writeln(" out.write(/"<attributes>\n/");");
   writeAttributes();
  myXMLFile.writeln(" out.write(/"</attributes>\n/");");
  myXMLFile.writeln(" out.write(/"</class>/");");
  myXMLFile.writeln("};");

  }
```

```
   /*********************************************************
******************/
  /**
   * Write-out attributes, operations, associations, etc. of a class.
   */
  function writeAttributes(xmlClassObj)
   {
    featureVector = xmlClassObj.getFeatureList(null);

   for (i=0;i<featureVector.size();i++)
   {
     thisFeature = featureVector.elementAt(i);
     thisFeatureType = new
java.lang.String(thisFeature.getClass().getName());
     if (thisFeatureType.equals("mil.darpa.ngii.uml.Attribute"))
     {
     myXMLFile.writeln("out.write(/"<attribute>\n/");");

myXMLFile.writeln("out.write(/"<name>"+thisFeature.getName()+"
</name>\n/");");

myXMLFile.writeln("out.write(/"<type>"+thisFeature.getType().
getName()+"</type>\n/");");

myXMLFile.writeln("out.write(/"<value>/"+"+thisFeature.getName()+"
+/"</value>\n/");");
       myXMLFile.writeln("out.write(/"</attribute>/");");
      }
    }
  }
```

This portion of template code when applied to a simple class:

Class A with attributes:

String name
String address
long age

would produce the following code:

```
public void writeToXML(StringWriter out)
{
 out.write("<class>");
 out.write("<classname>A</classname>");
 out.write("<attributes>");
out.write("<attribute>");
out.write("<name>name</name>");
out.write("<type>String</ type >");
out.write("<value>"+name+"</ value >");
out.write("</attribute>");
out.write("<attribute>");
out.write("<name>address</name>");
out.write("<type>String</ type >");
out.write("<value>"+ address +"</ value >");
out.write("</attribute>");
out.write("<attribute>");
out.write("<name>age</name>");
out.write("<type>long</ type >");
out.write("<value>"+age+"</ value >");
out.write("</attribute>");
 out.write("</attributes >");
out.write("</class>");
};
```

Service Composition

Service composition is the second area we focused on, and it proved to be the most challenging. Composing components within a system is usually a process of plugging in interfaces to well-defined units of functionality, such as Java Beans. Composition of services within an object in a systems schema is much more difficult. We discovered and implemented a number of different ways to compose services without affecting other aspects of the objects although each comes with its own unique issues. The first approach we took was to have the template developer insert calls to outside functions/methods at the correct place in the generation process. This approach, while it did work, did not prove to be very scaleable to a large number of different services because of the knowledge required about each service by the template developer. The second approach was to allow a template developer to implement a set of interfaces, which get calls based on the type of interface or based on template execution. This approach proved to be much more scaleable to a wide number of optional services, but does require the template developer to be much more versed in software development because it currently works only with the Java templates.

Template Modification and Addition

Our third area of focus was the ease of extending and adding new templates. Templates can currently be written in either Java or ECMAScript. Java templates allow for many developers to use the same language that they are using to code their templates. ECMAScript allows developers who have used VBScript or JavaScript to jump in and begin making use of a powerful development tool.

Our conclusion from our work with the code generation template was to concentrate on the Java-based templates. This conclusion was reached based on having the power of a full object-oriented programming language and using the language most developers were familiar with. In addition, because experts in the different areas of software development are usually the people writing templates, they prefer to write in a language that they commonly use.

BENEFITS

Many of the benefits of code generation are obvious, such as the decrease in time to market of new applications and systems, reduction in the amount of new code to be tested, and a reduction in the number of human errors. In this section, we will explore time reduction and some of the other benefits of code generation.

Code generation allows reuse of one of the scarcest resources in most companies: specialized experts. Experts in distributed transactions, security, or concurrence can be used to write specialized templates, thus allowing for corporate capture of that specialized knowledge and providing a force multiplier to other developers in an organization. Code generation also allows groups to define how they want the code to "look." Styles and enterprise-wide coding standards can be enforced by using templates that follow the standards. Because Quava allows the user to select which sets of templates to apply to their model, developers can experiment with a wide array of architectures and design patterns to see which best fits their specific requirements. Finally, code generation allows developers to be free of their underlying technology. Currently, when a new technology comes out, the developer must go back and re-code an application or system to make use of it. With code generation, new technologies can be merged with current systems, or underlying technologies can be completely replaced by new technology.

Reduction in Development Time

A reduction in development time is the main reason for using code generation techniques. Quava allows the developer to jump straight from the design into the coding phase with very little effort. Normally, the developer is handed a design document and must start from, at best case, generated code skeleton, or at worst case, from scratch. Quava reduces the amount of code a developer must write far more than generators that provide a code skeleton because it is generating object behavior, not just code file structure. Take the example used above for an XML streaming method. This would not be hard to write by hand, but why waste the developer's time doing something that could be generated? A reoccurring benefit of generating methods such as the XML streaming is that any time the model changes, those changes are quickly reflected in the source code. Eliminating human errors that result from typos and simple logic errors also reduces development time. Once a template has been tested, the code that it produces requires far less code testing, allowing the tester to focus more on the business logic of the system.

In our research on code generation, we measured a number of projects with varying object schemas to gather some quantifiable numbers of the kinds of savings code generation could produce. Table 1 shows values captured from some of these projects. The values for lines of code generated have been rounded off to the nearest thousand.

Overall, code generation has been proven to increase the speed at which systems and applications can be implemented, and, with Quava's generation technologies, the reduction is magnified by the experience of the developer.

TABLE 1. Code generation case study.

Case	Number of Classes	Average Attributes per Class	Average Operations per Class	Lines of Code Generated
A	7	30	24	8,000
B	120	22	6	257,000
C	321	12	8	750,000

Force Multiplier

Concurrency issues, complex services issues, and other difficult programming tasks can be encapsulated in templates. By having your best software engineers develop templates, every software engineer that generates an object server with that template may take advantage of their knowledge. In essence, with a software development model where experts create templates and junior programmers develop applications using object servers generated from such templates, an organization can produce much more high-end software. Of course, the exact value to the organization is only measurable by the number of times a template can be used.

Standardization of Enterprise Rules

By code generating the entire set of object servers with the same templates, a system engineer is guaranteed adherence to these enterprise rules. Different developers can interpret enterprise rules differently. Ambiguities in the software requirements specification can lead to major additional costs later on in the software development process [10].

If developers are allowed to produce object servers with different tools or different templates, it is impossible to guarantee that the system will perform as intended. These differences may even allow for correct execution when the interpretation is constant throughout the enterprise. However, when these different interpretations exist in the same enterprise, errors occur. When the problem domain consists of millions of objects and thousands of object servers, the only feasible solution is to code generate the object servers.

Experimentation

By allowing a system engineer to try different service implementations and middleware without having to encode all of the possible combinations by hand, a system engineer can develop prototypes of multiple test architectures and evaluate their characteristics in realistic deployment environments.

One DARPA project ran into trouble when the deployment environment proved to be less reliable than it was assumed to be. The project used hand-held computers networked with radio waves. When the connections between the hand-held computers proved unreliable, the system performance was severely impacted. Basically, the system would connect to the object servers only to be disconnected by unreliable communications within minutes. The system spent most of its resources establishing and re-establishing connections. The project was able to move from a connection-based architecture using CORBA to a connectionless architecture using HyperText Transfer Protocol (HTTP)/XML by regeneration of the object servers with different templates.

Technology Evolution

By building your data access and dissemination layer for the enterprise with Quava, your enterprise architecture can handle changes in software technology more readily. When advanced implementations of core services become available, a new template that implements the glue code between the new service implementation and the objects is created, and the object servers are regenerated without having to change any client application software. Also, when new middleware technologies arise, the object servers can be regenerated with additional interfaces so that the object servers can support client applications using the previous interfaces and new client applications using the new interface simultaneously. Older interfaces can be removed when client applications no longer need them by regenerating the object servers without the deprecated interface.

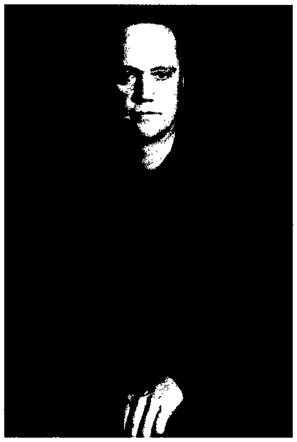
CONCLUSION

In our research, we found that model-driven code generation was a very promising technology with many benefits to the software practitioner. The benefits of using this approach in an enterprise help elevate many of the more substantial problems faced when developing large-scale systems. The openness and flexibility of the Quava implementation gives great support to life-cycle maintenance and software evolution of the system.

REFERENCES

1. Siska, C. 1998. "A Processor Description Language Supporting Retargetable Multi-Pipeline DSP Program Development Tools," *Proceedings on 11th International Symposium on System Synthesis*, 2–4 December, Taiwan, China, pp. 31–36
2. Bringmann, O., W. Rosenstiel, and D. Reichardt. 1998. "Synchronization Detection for Multi-Process Hierarchical Synthesis," *Proceedings on 11th International Symposium on System Synthesis*, 2–4 December, Taiwan, China, pp. 105–110.
3. Leone, M. and P. Lee. 1994. "Lightweight Run-Time Code Generation," *Proceedings of the ACM SIGPLAN Workshop on Partial Evaluation and Semantics-Based Program Manipulation*, June.
4. Engler, D. 1996. "VCODE: A Retargetable, Extensible, Very Fast Dynamic Code Generation System," *Proceedings of the 23rd Annual ACM Conference on Programming Language Design and Implementation*, 21–24 May, Philadelphia, PA, pp. 160–170.
5. Berzins, V., O. Ibrahim, and Luqi. 1997. "A Requirements Evolution Model for Computer-Aided Prototyping," *Proceedings of the 9th International Conference on Software Engineering and Knowledge Engineering*, Madrid, Spain, June.
6. Shing, M., V. Berzins, and Luqi. 1996. "Computer-Aided Prototyping System (CAPS)," *Proceedings of the Software Technology Conference*, Salt Lake City, UT, April.
7. Attie, P. and E. Emerson. 1989. "Synthesis of Concurrent Systems with Many Similar Processes," *Proceedings of the 16th Annual ACM Symposium on Principles of Programming Languages*, 11–13 January, Austin, TX, pp. 191–201.
8. Lewis, J., S. Henry, D. Kafura, and R. Schulman. 1991. "An Empirical Study of the Object-Oriented Paradigm and Software Reuse," *Conference Proceedings on Object-Oriented Programming Systems, Languages, and Applications*, 6–11 October, Phoenix, AZ, pp. 184–196.
9. Bhansali, S. 1995. "A Hybrid Approach to Software Reuse," *Proceedings of the 17th International Conference on Software Engineering Symposium on Software Reusability*, 29–30 April, Seattle, WA, pp. 215–218.
10. Henderson-Sellers, B. and J. Edwards. 1990. "Object-Oriented Systems Life Cycle," *Communications of the ACM*, vol. 33, no. 9, pp. 142–159.

❖



William J. Ray
MS in Software Engineering,
Naval Postgraduate School,
1997
Current Research: Enterprise architectures; distributed systems; object-oriented technologies.

Andy Farrar
BS in Computer Science,
San Diego State University,
1992
Current Research: Middleware technologies; software synthesis; distributed systems.

CINC 21 Advanced Concept Technology Demonstration

Richard N. Griffin
SSC San Diego

ABSTRACT

This paper describes an Advanced Concept Technology Demonstration (ACTD) entitled Commander-in-Chief for the 21st Century (CINC 21) and documents the involvement of SSC San Diego personnel in the ACTD. The goal of the ACTD is to create a highly visual, dynamically updated capability to develop and understand the CINC's theater situation, plans, and execution status during multiple, simultaneous crises involving joint, coalition, and humanitarian agencies based on shared knowledge and collaboration across secure and optimized networks. The paper describes operational needs and focuses on the application of technologies in specific areas. CINC 21 is a 5-year program consisting of 3 years of "development and integration" and 2 years of "residual support and transition."

OVERVIEW

Commander-in-Chief for the 21st Century (CINC 21) was a Fiscal Year 2000 (FY 00) new-start Advanced Concept Technology Demonstration (ACTD). The Joint Requirements Oversight Council (JROC) approved CINC 21 on 11 February 2000, and Congressional approval followed on 13 March 2000. The program consists of 3 years of development and integration and 2 years of residual support and transition.

CINC 21's mission is to develop and assess new command and control concepts for improving the speed and effectiveness of joint, coalition, and inter-agency operations by leveraging advances in visualization, knowledge management, information management, and network technologies.

CINC 21 directly addresses the emphasis that Joint Vision 2010 (JV 2010) places on Information Superiority and Decision Superiority. JV 2010 describes Information Superiority as the ability to collect, process, and disseminate an uninterrupted flow of information while exploiting or denying adversaries' ability to do the same. However, Information Superiority, while necessary, is not sufficient. Success in any operation requires the ability to effectively use and quickly exploit information. JV 2020 refers to this ability as Decision Superiority and states "... creation of information superiority is not an end in itself...we have a competitive advantage only when it is effectively translated into superior knowledge and decisions."

Specifically, CINC 21 will address Decision Superiority by: improving the ability of the CINC's "extended" staff to track and manage multiple, simultaneous crises; enabling synchronized understanding of operations between CINCs and the commanders of joint task forces; instituting enhancements to the information infrastructure to match operational needs (including coalition and interagency needs); and increasing the speed of command decision-making to gain and maintain the strategic advantage. Table 1 lists the lead organizations executing CINC 21.

The CINC 21 team concluded the first year of the development program in October 2000 by conducting a successful major demonstration at United States Pacific Command (USPACOM) headquarters. The primary audience for the demonstration included the directors of each staff component; the Deputy CINC, LT Gen Case; and the Deputy Chief of Staff, MG Lowe. Based on the success of this demonstration, the development team adopted a model of delivering technology, configured in operational packages, in a development cycle of 4-month increments. Spiral I was approved in February 2001, with delivery of the technology scheduled for mid-May 2001. The first major Military Utility Assessment (MUA) opportunity for CINC 21 ACTD delivered under the spiral development process will take place during Exercise KERNEL BLITZ (Experimental) (KB (X)) in June 2001. Subsequent development spiral deliveries will occur in September 2001, January 2002, and May 2002. Other MUA events are currently undetermined, but the culminating "graduation event" will be scheduled for a USPACOM Exercise in the fourth quarter of FY 02.

TABLE 1. Participating organizations.

Deputy Under Secretary of Defense for Advanced Systems and Concepts (DUSD [AS & C])	ACTD Oversight Dr. Robert Popp
U.S. Pacific Command (USPACOM)	Operational Manager Mr. Randall Cieslak
Office of Naval Research	Technical Manager Dr. Sue Hearold
Defense Information Systems Agency	Deputy Technical Manager LTC Riki Barbour
Space and Naval Warfare Systems Command (SPAWAR)	Transition Manager Mr. John Quintana

FY 03 and FY 04 are transition years consisting of three major activities:

1. Providing operations and maintenance (O&M) support for leave behind/residual capabilities,
2. Continuing transition planning with acquisition sponsors and programs of record,
3. Continuing assessments for the Defense Information Infrastructure Common Operating Environment (DII COE) and modifying applications as necessary to meet compliance requirements.

Objectives

Detailed objectives for the ACTD are stated in the CINC 21 Implementation Directive. U.S. Commander in Chief, U.S. Pacific Command (USCINCPAC) defined Critical Operational Issues (COIs). Table 2 shows the relationship between the COIs and the CINC 21 objectives.

TABLE 2. Objectives and critical operational issues.

Objectives	Critical Operational Issues
· Improve situational awareness and understanding through a) shared understanding of operational situation, b) scaleable and tailorable visualization, c) advanced decision support and knowledge management tools.	· Can advanced visualization technology empower individuals to process, digest, and assimilate large volumes of information, thereby enabling faster, more effective decisions? · Can knowledge management technology integrate information, context, and rules to increase understanding and, therefore, improve decision-making?
· Demonstrate and synchronize distributed decision-making, collaboration, and information management/information dissemination tools among joint, coalition, inter-agency, and non-governmental organization partners.	· Can collaboration tools be used to overcome the tyrannies of time, distance, and system disparity?
· Enable command of the information enterprise through advanced enterprise management tools and user-specified and prioritized operational products.	· Can the collection of networks, databases, and applications be enhanced to optimize the flow of information, with security assurance, across multiple network enclaves?

Concept of Operations (CONOPS)

At the center of CINC 21's Concept of Operations (CONOPS) is a knowledge-enabled information sphere with tools and applications that will improve situational awareness and understanding, provide the ability to collaborate as necessary, and manage the information enterprise while transforming and accelerating the decision processes that support the management of crisis-contingency operations, the CINC's theater engagement policy, and supporting staff processes.

The CONOPS for crisis operations includes expanding the ability of warfighting CINCs to handle multiple crises by delegating planning and execution to distributed crisis management cells and by simplifying the information flow to CINC and Commander, Joint Task Force (CJTF) decision cells. A combination of intelligent information management and continuous collaboration with multiple crisis cells will accomplish this task. Benefits will accrue to the CINC headquarters, supporting and supported CINCs, subordinate unified commands, Department of Defense (DoD) and non-DoD agencies, non-government organizations, and coalition partners. CINC 21 addresses the need for CINCs and CJTFs to operate in this complex world environment by exploiting the power of visualization to convey knowledge and understanding.

In addition to traditional military operations, the 21st century environment makes it necessary to participate in a wide variety of theater engagement activities. These mission areas, sometimes known as Military Operations-Other-Than-War (MOOTW), include refugee control, humanitarian assistance, disaster relief, non-combatant evacuation, public security/law and order, support to host governments, mediation/negotiations, and demilitarization operations. All these operations put a premium on open-source/unclassified information that can be readily shared with all participants.

The desired outcome for this environment is threefold: (1) the necessary mature and maturing tools will be integrated to enable open-source information to be added to the information-gathering systems available to the USPACOM virtual staff, (2) the information will be compatible with decision-support software tools that enable assessment, evaluation, and prioritization of appropriate courses of action (COAs), and (3) the open-source information should be accessible from a mobile or remote command site/"cell."

As Figure 1 shows, CINC 21 will provide a highly visual, dynamically updated capability to develop and understand the CINC's theater situation, plans, and execution status during multiple, simultaneous crises involving joint, coalition, and humanitarian agencies based on shared knowledge and collaboration across secure and optimized networks. CINC 21 will provide the following capabilities:

- User-tailorable, integrated situation display
- Enhanced visualization of information so decision-makers can quickly interpret, assimilate, and act
- Secure access to relevant information at its source on demand (demand can be from user or intelligent agent)
- Distributed collaborative environment enabling rapid command and control and access to expertise at its source—collaboration as a basic service
- Enhanced security by providing capability to establish trusted network relationships on demand
- Information flow monitoring and dynamic allocation of resources to optimize distribution of information based on commanders' priorities

Technical Approach

CINC 21 seeks to provide an enhanced decision support environment for the CINC and its extended staff through mature commercial and government software packages. The objectives of the CINC 21 ACTD can be mapped into the four technical areas listed in Table 3.

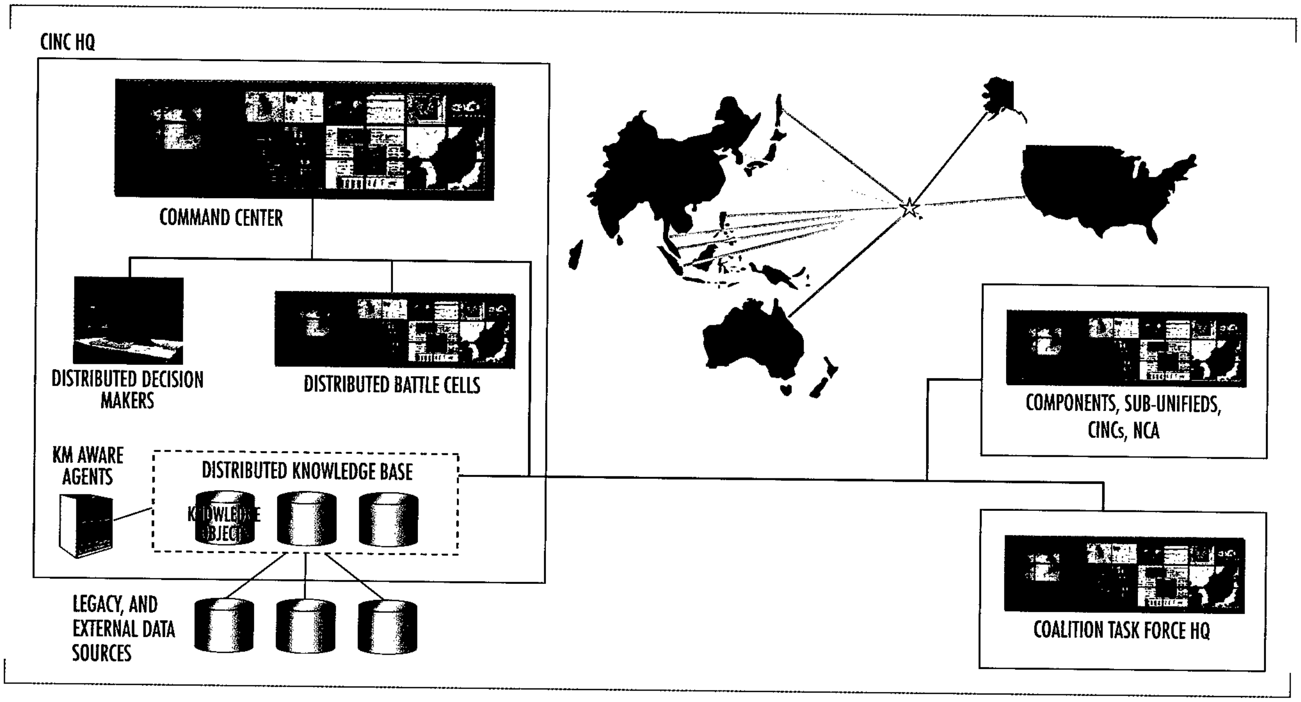

CINC HQ
COMMAND CENTER
DISTRIBUTED DECISION MAKERS
DISTRIBUTED BATTLE CELLS
KM AWARE AGENTS
DISTRIBUTED KNOWLEDGE BASE
KNOWLEDGE OBJECTS
LEGACY, AND EXTERNAL DATA SOURCES
COMPONENTS, SUB-UNIFIEDS, CINCs, NCA
COALITION TASK FORCE HQ

FIGURE 1. CINC 21 operational concept.

System Design, Engineering, and Integration

The CINC 21 responsibility for system design, engineering, and integration is assigned to the System Development Team led by Ray Glass (SSC San Diego). This team is responsible for all development activities leading up to the hand-off of a robust, configuration-managed hardware/software solution to the CINC 21 Implementation Team, led by Tom Tiernan, SSC San Diego.

TABLE 3. CINC 21 technical areas.

Technical Area	Description
Data interoperability (information management, knowledge management, network infrastructure)	Provide improved mechanisms for sharing information across the CINC's staff and to enable more effective and efficient production of cross-staff decision products.
Information infrastructure enhancements (information management, knowledge management)	Provide upgrades to the CINC's information infrastructure that improve decision-making, foster greater inter-agency and coalition interaction, and improve security.
Knowledge wall environment (visualization)	Provide a structured environment that allows the rapid development and easy sharing of a wide range of correlation, visualization, and collaboration services. As an adjunct to this activity, CINC 21 will pursue the delivery of multi-panel desktop and wall-based displays as residual capabilities.
Operational packages (knowledge management)	Develop specific operational capabilities for USPACOM and United States Strategic Command (USSTRATCOM) by using a complete set of Extensible Markup Language (XML), Decision Tagged Data (DTD), databases, correlation, and visualization plug-ins to create useful end-to-end services.

The breakdown of the CINC 21 development activities into the four areas has been done to carefully split the responsibilities of the Development Team so that they can concentrate more fully on their primary objectives. Designers of the general framework services will not be inclined to shortcuts because of pressure in delivering specific operational packages. Operational package developers will be freed from the responsibilities of building and maintaining the core services. Integration, test, and configuration management has been separated from both activities to ensure unbiased independent verification and validation (IV&V). Finally, the operational support activity has been called out separately to ensure that support to the Implementation Team does not have a resource impact on other system development activities.

To ensure a common foundation for all three classes of CINC 21 users, the system development activity will be divided into three parts: a system design effort that will design and develop the user-independent CINC 21 foundation, an operational package development and operational support activity effort that will provide domain-specific products to operational users, and an implementation management effort responsible for integrating the operational packages into operational use. Figure 2 shows the system development approach.

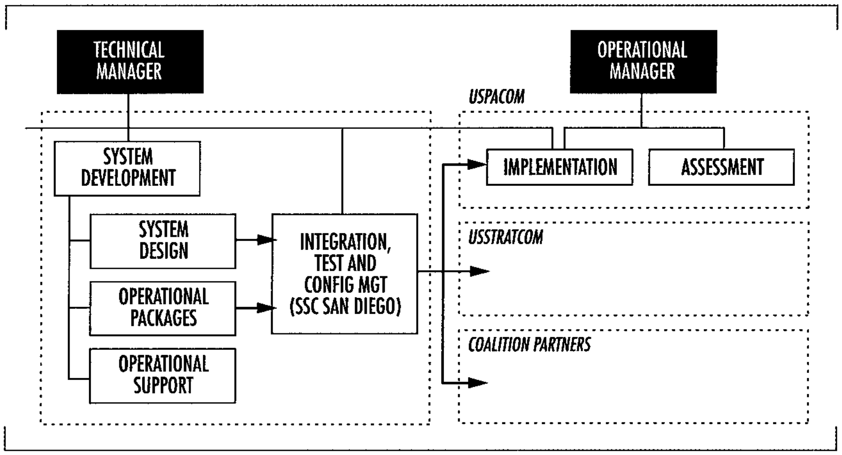


FIGURE 2. System development approach.

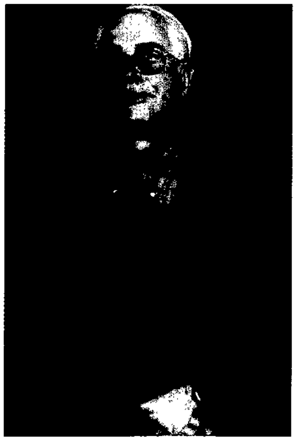
CONCLUSION

As stated in the introduction to this paper, CINC 21's objective is to increase the speed of command across the spectrum of operations by controlling and exploiting an information-rich environment. This objective demands advanced technologies linked to advanced concepts. Within CINC 21, we are exploring concepts and technologies that not only improve the ability to collect, process, and disseminate information, but also fundamentally change the way warfighters use that information by applying tools and processes that create knowledge and understanding. Today, we drown people in information, but leave them starving for knowledge. With CINC 21, we will show how we can significantly improve the ability to command and control forces by providing a more visual, structured, and interactive command environment.

ACKNOWLEDGMENTS

SSC San Diego personnel played key roles in the development of operational requirements for advanced technology and in shepherding the ACTD proposal through the approval process. They continue to play major roles in the management of the ACTD and in the integration of technological solutions to apply to warfighter requirements. Jeff Grossman, Tom Tiernan, and Dick Griffin were major players in the development of concepts and requirements. Sue Hearold (on loan to the Office of Naval Research [ONR]) is the Technical Manager, Ray Glass is the System Development Manager, Tom Tiernan is Systems Implementer, Pete Wussow leads the development of decision-focused visualization tools, and Dick Griffin is Deputy Operational Manager and leads the military utility assessments phase. Mike Reilley, USCINCPAC Science and Technology Advisor, provides oversight and liaison with technology developers.

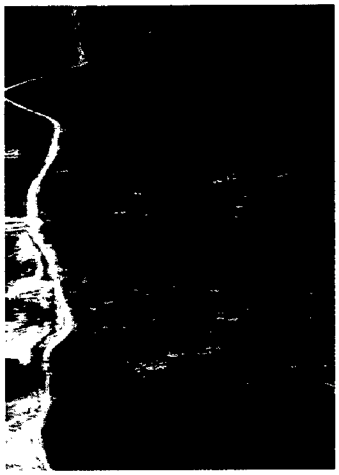
❖



Richard N. Griffin
MA in International Studies, The Johns Hopkins University, 1977
Current Work: Deputy CINC 21 Operational Manager, U.S. Pacific Command.

2

Data Acquisition and Exploitation



2 DATA ACQUISITION AND EXPLOITATION

Evolutionary Control of an Autonomous Field

Mark W. Owen
SSC San Diego

Dale M. Klamer and Barbara Dean
Orincon Corporation

ABSTRACT

An autonomous field of sensor nodes must acquire and track targets of interest traversing the field. Small detection ranges limit the detectability of the field. As detections occur in the field, detections are transmitted acoustically to a master node. Both detection processing and acoustic communication drain a node's power source. To maximize field life, an approach must be developed to control processes carried out in the field. This paper presents an adaptive threshold control scheme that minimizes power consumption while still maintaining the field-level probability of detection. The power consumption of the field of sensor nodes is driven by the false alarm rate and target detection rate at the individual sensor nodes in this problem formulation. The control law to be developed is based on a stochastic optimization technique known as evolutionary programming. Results show that by dynamically adjusting sensor thresholds and routing structures, the controlled field will have twice the life of the fixed field.

INTRODUCTION

The Office of Naval Research (ONR) established the Deployable Autonomous Distributed System (DADS) program (Figure 1) to demonstrate the feasibility of increased performance for an advanced tactical/surveillance system that operates as a field of underwater distributed sensor nodes. The goal of DADS is to demonstrate the feasibility of a cooperative field-level detection and data fusion system that increases performance at a reduced cost. Given limited power, the objectives are to use distributed detection and data fusion to increase the lifetime of the field (reduced power consumption), decrease the false alarm rate of the field over that of the individual nodes, increase the field-level detection, increase the probability of correct classification, and increase the accuracy of target position estimates [1, 2, and 3].

A DADS field consists of individual sensor nodes operating autonomously. Each sensor node uses a set of acoustic and electromagnetic sensors to provide coverage of a small area of interest. Each DADS sensor node uses a matched-field tracking algorithm to provide target detections consisting of position, velocity, and classification information. Once a detection is constructed at a sensor node, the data are transferred to a DADS master node where field-level data fusion is performed.

Detection Theory

In the DADS program, a need exists to identify what constitutes target detections from the field of autonomous sensor nodes. The DADS program also requires an optimization algorithm to route communication messages efficiently, using as little power as possible. A field-level control/detection scheme is sought to detect targets of interest at a given field-level probability and to route messages optimally by using a minimal amount of power. Control of an autonomous set of sensor nodes is needed to meet a desired probability of detection for the field and to extend the life of the field.

To construct a field-level detection, we now define what is required to call out a field-level detection. Each sensor node contains an acoustic sensor suite and an electromagnetic sensor suite. To report a detection, both the acoustic and magnetic sensors must detect a target at a sensor node. Once one node has detected the target, a second node nearby is cued and another sensor node must detect the target. Once this second sensor node detects and reports the target, a field-level detection is called and reported

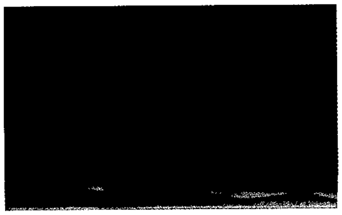


FIGURE 1. Field of DADS autonomous sensor nodes.

out by the master node for field-level fusion. Each sensor node has a threshold for the sensor suite given by an operating point on a receiver operating characteristic (ROC) curve as shown in Figure 2. The operating points on the figure are labeled R1 and R2 and represent different signal-to-noise ratio (SNR) levels for the sensor suite. Choosing different operating points on the ROC curve yields different probabilities of detection and probabilities of false alarm. A constant field-level probability of detection is desired for operation of the field of sensor nodes. By adjusting threshold levels at the sensor suite, that is, moving up and down operating points on the ROC curve at each sensor node, a constant field-level probability can be achieved.

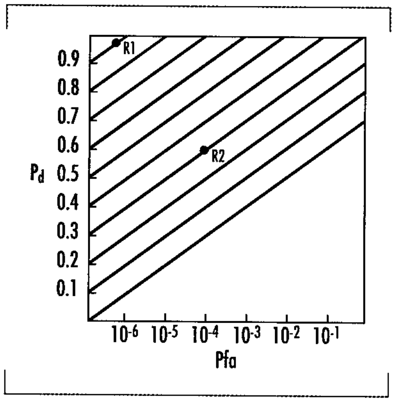


FIGURE 2. Typical ROC curve.

Besides controlling the thresholds at the individual sensor suites at each node, another problem is to minimize the power consumption of the individual sensor nodes while meeting the field-level probability constraint. This issue addresses the routing of communication messages through the distributed field of sensor nodes. As messages are passed from sensor node to sensor node and finally arrive at the master node, the battery level is drained by the amount of communication power spent transmitting and relaying detections acoustically.

A field-level controller will adjust the detection threshold levels at each sensor node to meet the desired field-level probability of detection and to perform optimal routing of messages through the field. A typical example of a point on a ROC curve is shown in Figure 3.

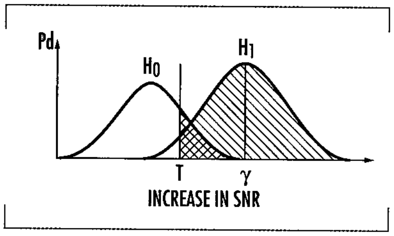


FIGURE 3. A single point from a ROC curve.

A brief overview of detection theory is provided below [4]. In Figure 3, two possible hypotheses, labeled H0 and H1, are shown. H0 is the false alarm hypothesis and H1 is the detection hypothesis. The threshold T is used to determine whether or not the SNR is high enough to call out a detection. The SNR in the figure is labeled γ. Under the two Gaussian curves, a probability of detection and a probability of false alarm can be determined. Integrating the H0 probability density function (pdf) from T to ∞, the false alarm probability is calculated. Integrating the H1 pdf from T to ∞, the probability of detection is calculated. Figure 4 shows several SNRs from a chosen ROC curve operating point. The objective of the field-level controller is to adapt the sensor node thresholds to acquire a target of interest and detect it successfully through the field. In the figure, the graph labeled nominal is shown to demonstrate a chosen operating point for the sensor node. The next two graphs show a decrease in SNR and an increase in SNR, respectively. As SNR levels vary, a target may become easier or more difficult to detect although the probability of false alarm remains constant across all three graphs. Only the probability of detection decreases or increases due to the SNR of the target. Our task is to adjust thresholds dynamically to make sure the target is acquired and tracked as it passes through the field. To do this, we will lower thresholds for subsequent cued detections to increase the detection range at a sensor node, but at the same time we increase the number of false alarms from a sensor node. When adjusting these thresholds at each sensor node, we must maintain a constant field-level probability of detection. A simple example of this threshold adjustment is to use a bathtub analogy. If one side of the bathtub water is pushed down, water on the other side of the tub will rise. This example shows what we will do when adapting thresholds: we will lower a certain set of sensor node thresholds while raising another set.

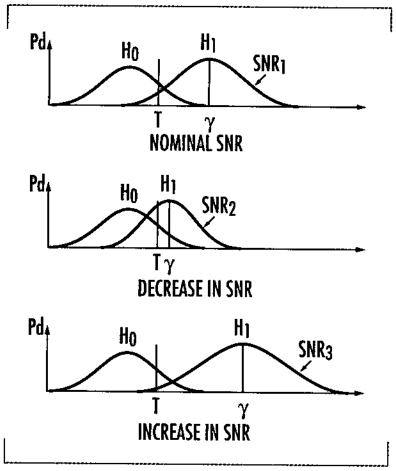


FIGURE 4. Possible detection curves.

Threshold Adaptation

Figure 5 shows a cookie cutter example of a field of sensor nodes. Each sensor node has a defined detection range given in red (small circles) for a high threshold (low false alarm rate, high SNR) and another detection range shown in blue (large circles) for a low threshold (high false alarm rate, low SNR). This figure demonstrates the adaptive process that must occur for the DADS field of sensor nodes to detect and continue to detect a target as it passes through the field.

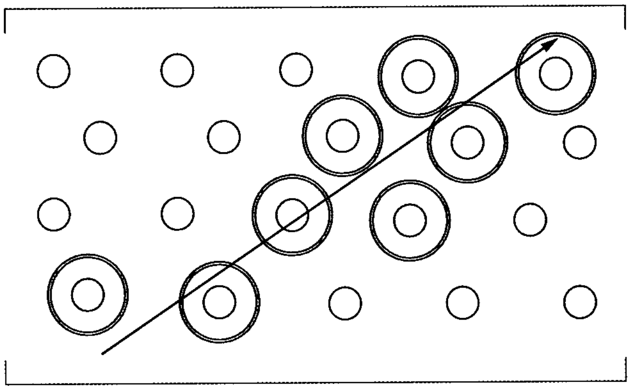


FIGURE 5. Sensor node threshold adjustments via field-level control.

If the field were static, the small red circles would dictate the area of coverage in which the field could pick up detectable targets. In the figure, a hypothetical target has been drawn by a black line with an arrow at the tip. If the threshold were held at this higher level, only one possible detection might occur as this target traversed the field of sensor nodes. By lowering the thresholds (larger blue circles), which is done by cueing the field, a broader coverage of the field is achieved. The figure shows that up to four possible detections on a target of interest can occur by lowering the sensor node thresholds. This improved detectability concept will improve the overall field-level data fusion by providing more contact information than previously capable with a static set of sensor node thresholds. By lowering the threshold though, a larger number of false alarms can occur and cause power to be drained from the sensor nodes. False alarms also make the data fusion problem at the master node more susceptible to miscorrelation. Therefore, dropping all of the sensor node thresholds is not acceptable because it will limit the system operation. As explained previously, we will lower thresholds and raise thresholds at individual sensor nodes to maintain the desired field-level probability of detection while maximizing the life of the field.

2-of-2 Field Detector

To adjust thresholds, we propose to use a baseline model of a 2-of-2 detector. The detector will use communication costs, probabilities of detection and false alarm, node spacing of the field, and signal processing parameters used at the sensor node sensor suite. This formulation shows that false alarms as well as target detections drain the power at each sensor node. We will now present our baseline model equation for field-level control as derived in [5]. This formulation will allow the complete field to be controlled by the master node in the DADS system. The baseline model equation is as follows. The estimated power $\hat{P}^{(n)}$ consumed over a period of time T at each node n, $n = 1,\dots, N$, is given by

$$\begin{aligned}\hat{P}^{(n)}(T) = &\sum_{k=1}^{\rho_s T} C_{on} + [1-(1-F_1^{(n)}F_2^{(n)})^{N_p}]C_k^{(n)} \\ &+\sum_{n' \varepsilon R_k^{(n)}} [1-(1-F_1^{(n')}F_2^{(n')})^{N_p}]C_k^{(n)} \\ &+\sum_{n' \varepsilon B_k^{(n)}} [1-(1-F_1^{(n')}F_2^{(n')})^{N_p}]\left[1-(1-F_1^{(n)}F_2^{(n)})^{\rho_s \delta N_p P[1+sD^2]/(\pi(r_d^{(2)})^2)}\right]C_k^{(n)} \\ &+\sum_{n' \varepsilon R_k^{(n)}} \sum_{n'' \varepsilon B_k^{(n)}} [1-(1-F_1^{(n'')}F_2^{(n'')})^{N_p}]\left[1-(1-F_1^{(n')}F_2^{(n')})^{\rho_s \delta N_p P[1+sD^2]/(\pi(r_d^{(2)})^2)}\right]C_k^{(n)} \qquad (1)\end{aligned}$$

$$
\begin{aligned}
&+ \rho_T \mathrm{r_d}^{(n)}[1-(1-P_1^{(n)}P_2^{(n)})(1-F_1^{(n)}F_2^{(n)})^{N_p-1}]C_k^{(n)} / D \qquad \text{(1 contd)}\\
&+ \rho_T \sum_{n'\varepsilon R_k^{(n)}} \mathrm{r_d}^{(n)}[1-(1-P_1^{(n')}P_2^{(n')})(1-F_1^{(n')}F_2^{(n')})^{N_p-1}]C_k^{(n)} / D\\
&+ \rho_T \sum_{n'\varepsilon B_k^{(n)}} \mathrm{r_d}^{(n)}[1-(1-P_1^{(n')}P_2^{(n')})(1-F_1^{(n')}F_2^{(n')})^{N_p-1}]\\
&[1-(1-P_1^{(n)}P_2^{(n)})(1-F_1^{(n)}F_2^{(n)})^{[\rho_s \delta N_p P[1+sD^2]/(\pi(r_d^{(2)})^2)]-1}]C_k^{(n)} / D\\
&+ \rho_T \sum_{n'\varepsilon R_k^{(n)}} \sum_{n''\varepsilon B_k^{(n')}} [1-(1-P_1^{(n'')}P_2^{(n'')})(1-F_1^{(n'')}F_2^{(n'')})^{N_p-1}]\\
&[1-(1-P_1^{(n')}P_2^{(n')})(1-F_1^{(n')}F_2^{(n')})^{[\rho_s \delta N_p P[1+sD^2]/(\pi(r_d^{(2)})^2)]-1}]C_k^{(n)} / D
\end{aligned}
$$

where ρ_s is the basic sample rate and T is the time period of the estimated life of the node. The first term represents the power consumed C_{on} from the processor in the node. If the sensor node is on, a certain amount of processing power is drained from the battery. The second term represents the case that an initial false alarm is generated at node *n*, where $F_1^{(n)}$, $F_2^{(n)}$ are the probabilities of false alarm that are controlled by thresholds $T_1^{(n)}$ and $T_2^{(n)}$, and $C_k^{(n)}$ is the communication power used to transmit from node *n* to the next upstream node specified by the current communication route $R_k^{(n)}$ at time *k*. N_p is the size of the parameter space over which the detectors must test, e.g., if the detector must look over a discrete set of speed (say N_s) and closest point of approach (CPA), say N_{CPA} , thus giving $N_p = N_s N_{CPA}$. This is the second detection required for declaring a field-level detection from the field. The third term represents the case of a "downstream" node *n'* that generates a false alarm and node *n* is simply a passthrough; the communication route for node *n* at time *k* is specified by $R_k^{(n)}$. The fourth term represents the case that a false alarm is generated at node *n* as the result of being cued by another node *n'* in a set of neighboring nodes $B_k^{(n)}$. Specifically, *P* is the covariance of the track estimate at the time of the detection at the first node; $[1+sD^2]$ is the expansion factor for the track covariance until the second detection at the next node detection; $\pi(\mathrm{r_d}^{(2)})^2$ is the area of the detection space for the second sensor node; and *D* is the length of the sensor field. The fifth term represents the case of a downstream node *n'* that generates a false alarm as a result of being cued, and node *n* is simply a passthrough. The last four terms deal with the cases of a target present; ρ_T is the target rate. The sixth term represents a target detection at node *n*, where $P_1^{(n)}$, $P_2^{(n)}$ are the probabilities of detection, again controlled by the thresholds $T_1^{(n)}$ and $T_2^{(n)}$. This is a true target detection and not a false alarm. The seventh term represents the case of a downstream node *n'* detection where node *n* is simply a passthrough for the initial condition. The eighth term represents the case that a target detection is generated at node *n* as the result of being cued by another node *n'*. The final term represents a downstream node *n'* that generates a target detection as the result of being cued, and node *n* is simply a passthrough.

Given the current power $P^{(n)}$ available at each node, the estimated remaining power is

$$\varepsilon^{(n)}(T) = P^{(n)} - \hat{P}^{(n)}(T)\ .$$

The objective function for maximizing the life of the field is

$$\text{maximize } T,$$

subject to the constraints that each of the estimates of the remaining power is positive

$$\varepsilon^{(n)}(T) \geq 0\,,\ n = 1,\ldots, N$$

and the field-level probability of detection is specified by

$$\begin{aligned}\mathrm{PD} = {} & N(\varepsilon_1, \varepsilon_2,\ldots, \varepsilon_N)\pi(\mathrm{r}_\mathrm{d}^2)\left[1-\left(1-P_1^{(1)}P_2^{(1)}\right)\left(1-F_1^{(1)}F_2^{(1)}\right)^{N_p-1}\right]\\ & \times\left[1-\left(1-P_1^{(2)}P_2^{(2)}\right)\left(1-F_1^{(2)}F_2^{(2)}\right)^{\left[\rho\delta N_p P(1+sD^2)/\left(\pi r_d^{\,2}\right)\right]-1}\right]/A(D)\end{aligned}$$

where $N(\varepsilon_1, \varepsilon_2,\ldots, \varepsilon_N)$ is the number of nodes with nonzero power remaining and $\pi(\mathrm{r}_\mathrm{d}^2)/A(D)$ is the area covered by an individual node. The objective is to maximize field life T subject to meeting the field-level constraint by adjusting probability of detection / probability of false alarm threshold levels and varying communication routes (through $R_k^{(n)}$). By choosing appropriate thresholds at each sensor suite, the field-level probability of detection constraint can be met and the field life extended. An algorithm that will choose thresholds to meet the probability of detection constraint and extend the field life is discussed in the next section.

Evolutionary Programming

Evolutionary programming (EP) is a stochastic optimization technique applied in this paper to optimize routing of the sensor node message traffic at minimal power cost and to meet a field-level probability constraint. EP falls under the domain of Evolutionary Computation that contains other algorithmic techniques such as genetic algorithms (GAs), genetic programming, as well as others [6]. One of the main differences between EP and GAs is that EP performs a mutation operation while GAs perform a mutation operation and a crossover operation. Genetic algorithms also operate from the bottom up when finding a solution. EP is a top down approach to finding optimal solutions. An evolutionary algorithm is shown in Figure 6. In simple terms, an evolutionary algorithm starts out with a population of possible solutions to a problem. A population consists of parent solutions and their corresponding offspring solutions. This stochastic optimization technique allows the whole parameter space to be searched and evaluated for a best-fitting solution. In the figure, the initial solutions are called parents. Each parent solution can be a good first guess at the correct answer or a randomly chosen solution that may be very poor. Each parent has the ability to create a set of offspring solutions by mutation or by crossover if a genetic approach was used. Each parent solution is mutated by changing its state to form an offspring solution. This mutation can be Gaussian or some other linear or nonlinear deviation. Once the population of parents has been mutated and the offspring solutions are created, the population consisting of parents and offspring solutions is then scored, as shown in the figure. Scoring or evaluation of the population for our purpose is done to make sure the sensor nodes meet a defined field-level probability constraint with their defined threshold settings. A selection process is then performed whereby the next generation of parents are selected to evolve better and better solutions. This selection process chooses the solutions that passed the constraint in the scoring process by selecting the solutions that yield the largest amount of field life.

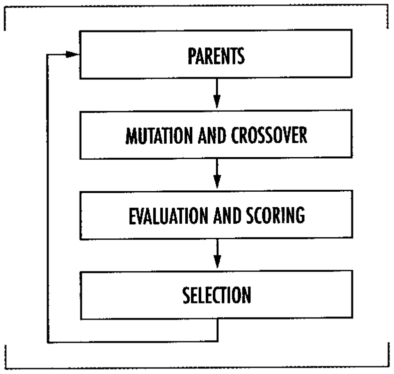


FIGURE 6. Evolutionary algorithm.

The standard EP approach consists of several steps (initialization, mutation, scoring, and selection) [6]. Initialization is performed by assigning thresholds to each sensor in the sensor suite (magnetic, acoustic) and using these thresholds, the sonar equation, and an error function to evaluate the probability of detection and probability of false alarm of the sensor node. This is done for each sensor node in the field given by

$$Pd(n) = 1/2{*}(1.0 - erf(T(n) - SL(n) + NL(n))) \quad (2)$$

and

$$Pd(n) = 1/2{*}(1.0 - erf(T(n) + NL(n))) \quad (3)$$

where Eq. (2) initializes the probability of detection Pd for sensor node n given its threshold T, the target source level SL, and the noise level at the sensor NL. Eq. (3) initializes the probability of false alarm Pfa for sensor node n given its threshold T, and the noise level at the sensor NL. This is performed for each sensor node until all thresholds and probabilities of detection and false alarm have been initialized. This fully initialized field of sensor nodes is deemed as a parent solution in the EP language and is a possible solution for the field-life problem. Possible solutions are defined as parents and are given as

$$P(k) = S(Pd(n), Pfa(n), T(n), R(n)) \quad (4)$$

where $P(k)$ are the k number of parents in the population solutions. Each solution S is made up of a field of sensor nodes with independent thresholds T, which dictate a Pd and Pfa for the sensor node, and a routing table R for communication with other nodes in the field. Once the population of parent solutions has been initialized, the EP algorithm is able to perform the next three steps (mutation, scoring, and selection) iteratively to converge to the best possible solution given time constraints and memory requirements of the system. The first step is the mutation process whereby parent solutions generate offspring solutions. Offspring solutions have the possibility of generating a better solution than their parents. This is the evolutionary step in the EP process. One of the mutation steps is to change the threshold at each sensor at a sensor node to yield a better solution. This is defined by

$$O[T(m,n)] = P[T(k,n)] + N(0,1) \quad (5)$$

where $O[T(m,n)]$ is the mutated threshold at offspring m for sensor node n, $P[T(k,n)]$ is the threshold at parent k for sensor node n, and $N(0,1)$ is a Gaussian random variable with zero mean and unit variance. Eq. (5) changes each parent's threshold to generate an offspring's threshold. Another mutation step is to change the routing table for communications at each node. This is defined by

$$O[R(m,n)] = P[R(k,n)] \pm Urv{*}c \quad (6)$$

where $O[R(m,n)]$ is the mutated communication routes at offspring m for sensor node n, $P[R(k,n)]$ is the communication routes at parent k for sensor node n, Urv is a Uniform random variable, and c is the number of possible nodes for sensor node n to communicate with. The number of communication routes can increase or decrease according to Eq. (6). Eq. (6) changes each parent's communication route to generate an offspring's communication route. Each parent can perform these mutation steps and generate as many offspring as desired. Once this is done, the new population of parents and offspring are scored and evaluated against the system constraints. For example, if the desired field-level probability of detection is 0.8, each solution is evaluated using

$$\begin{aligned}\mathrm{PD} = N\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)\pi(\mathrm{r_d}^2[1 - (1 - P_1^{(1)}P_2^{(1)})(1 - F_1^{(1)}F_2^{(1)})^{N_p - 1}] \\ \times[1 - (1 - P_1^{(2)}P_2^{(2)})(1 - F_1^{(2)}F_2^{(2)})^{[\rho\delta N_p P(1+sD^2)/(\pi r_d{}^2)] - 1}]/A(D) \quad (7)\end{aligned}$$

which is the probability of detection for a field of sensor nodes defined above. (See 2-of-2 Field Detector.) We will use a simulated annealing approach to meet this constraint. For example, if 0.8 is desired, we may allow solutions to lie between (0.7, 0.9) in the beginning and slowly converge toward 0.8 while we iterate. All solutions that pass this field-level probability constraint are then passed to the selection process. Selection is done by picking the best *k* solutions that meet the constraint and minimize the power consumption defined from the baseline model from Eq. (1). These best *k* solutions then become the parents for the next iteration. The process continues until the best solution is found. This evolutionary process extends the field life by optimizing the thresholds of the field and planning the optimal routes for message passing.

RESULTS

Now we present some results of our EP solution to the adaptive threshold control problem. These results are for a complete field of sensor nodes. Each node has a set of thresholds solved for by the EP algorithm as well as the optimal routes for communication to extend field life.

Simulation Overview

As stated previously, the claim of this paper is that it can be shown that field life can be doubled by using a field-level controller to dynamically adjust thresholds and routing structures, as compared to a fixed field that uses static thresholds and routing structures.

The EP software written for this paper generates solutions that are representative of a field under the control of a field-level controller. To make the comparison to a fixed field, a fixed-field implementation had to be generated.

The Fixed Field

The fixed field required a nominal routing structure and a set of sensor thresholds, which would meet the field-level probability of detection. To generate the nominal routing structures, a field initialization scheme was emulated. The emulation of this field initialization scheme consists of the following steps:

1. The Master Node broadcasts a Wakeup Message.
2. Any node that can hear responds with a Wakeup Response message. In this case, any node within the cookie cutter range can hear.
3. Nodes that responded to the Master Node will be direct communication routes. This means that these nodes will relay their packets directly to the master node.
4. Nodes that heard the Master Node will broadcast to their neighbors.
5. Any node that can hear within the cookie cutter range will respond.
6. If the node that responds does not have a destination node yet, the node that broadcast will become the destination node.
7. This sequence is repeated until every node in the field has been assigned exactly one destination node.

The above sequence generated a nominal routing structure for a fixed field as shown in Figure 7. In conjunction with the routing structures, sensor thresholds that met the field-level probability of detection were required. To obtain these thresholds, the EP model was run, and the thresholds from the optimal solution were used.

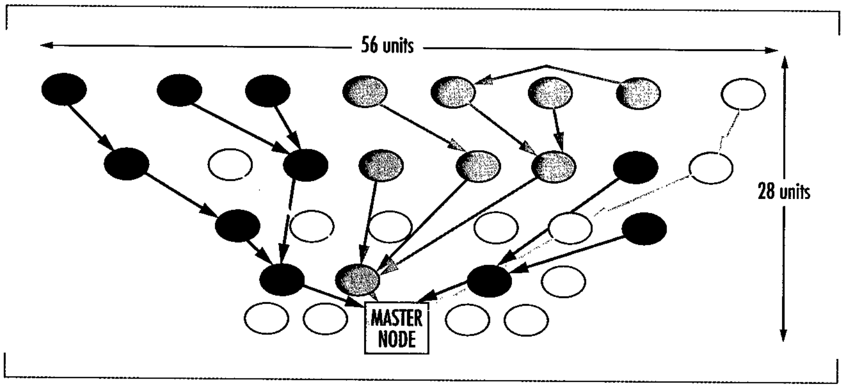


FIGURE 7. Fixed-field routes.

The Controlled Field

In the simulations, two types of results are generated for the controlled field. The first type is referred to as a "single optimized" solution. This solution is generated using the EP software. Once the EP algorithm finds an optimal combination of thresholds and routing structures, it uses that solution for the life of the field. Figure 8 shows the optimal routes found for the single optimized solution.

The second type of a controlled field solution is referred to as a "vector-optimized" solution. As with the single optimized solution, the EP algorithm finds a solution set, which maximizes field life. However, in this solution, the routes and thresholds can be adjusted every 24 hours, thus resulting in a vector of solutions. Because the control algorithm is run each day and the routes are potentially changed, it is not possible to show each daily graphical solution in this paper.

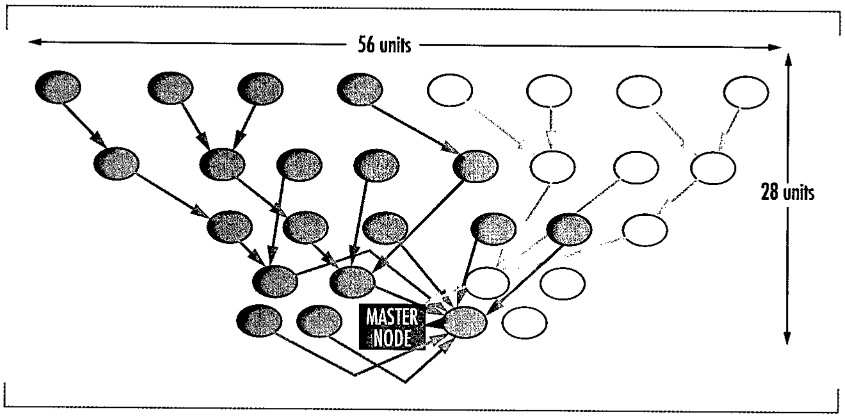


FIGURE 8. Single optimized field routes.

Field Laydown

Simulations were run for two field laydowns. In each laydown, the field consists of 30 sensor nodes and 1 master node arranged in a (56 by 28) unit grid. The difference between the two laydowns is the placement of the master node. In the first field laydown, the master node is a square box on the edge of the field as shown in Figures 7 and 8. In the second laydown, the master node is in the center of the field of sensor nodes.

Detector Types

The objective function defined previously (see 2-of-2 Field Detector) is for a 2-of-2 detector. This paper also defined an objective function for a 1-2 detector. The 1-2 detector requires an initial detection from the magnetic sensor on one node followed by a confirmed detection from the acoustic sensor on a second node. Results for both the 2-of-2 detector and the 1-2 detector are reported below.

Simulation Results

The results from the simulation are given in Table 1. The results are provided in units of days.

TABLE 1. Simulation results in days.

Field Laydown	Detector	Fixed Field	Single Optimized	Vector-Optimized
1	1-2	21	32	45
	2-of-2	40	70	118
2	1-2	26	45	55
	2-of-2	74	106	154

Figure 9 shows the results from running the fixed-field simulation. In the fixed field, the routing assignment was performed by using the minimum number of hops between the master node and each node in the field. This result is for the 2-of-2 detector processing for the second field laydown. It shows that running no optimization algorithm and just a greedy algorithm to assign a route for the field only yields a field life of 74 days. As shown in Figure 9, one single node begins to lose its power immediately. This node is the main communication node to the master node. Once one node in the field loses all of its power, the field is considered to be dead.

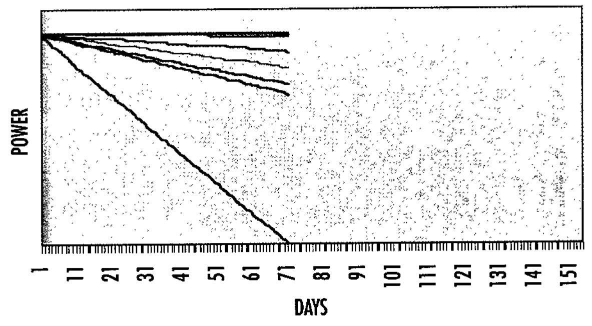


FIGURE 9. Fixed-field life.

Figure 10 shows the results from the single optimized field simulation. The routes for this result were calculated by running the EP algorithm once for the whole life of the field. This optimization result yielded a field life of 106 days for the 2-of-2 detector for the second field laydown. As shown in this figure, a single node still drives the field to death, but there are several other sensor nodes that are also losing power at a similar rate.

The field life was extended over the fixed-field implementation by using at least one planned optimal route for the whole simulation.

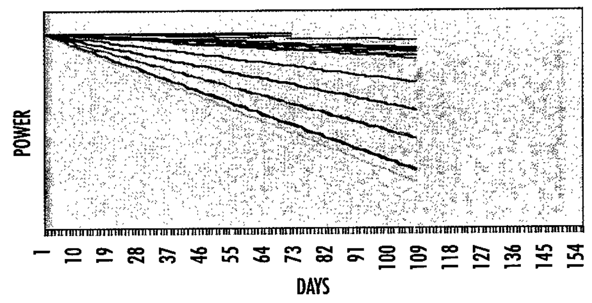


FIGURE 10 . Single optimized field life.

Figure 11 shows the results from the vector-optimized field simulation. This result has its routes recalculated each day by running the EP optimization algorithm. This optimization result yielded a field life of 154 days for the 2-of-2 detector for the second field laydown. As shown in this figure, a group of sensor nodes all lose power similarly at the same rate. Approximately one-third of the sensor nodes in the field died on day 154. This result more than doubled the life of the field over the fixed-field result of Figure 9. It also increased the life of the field from 106 days for the single optimized solution shown in Figure 10 to 154 days for the vector-optimized solution.

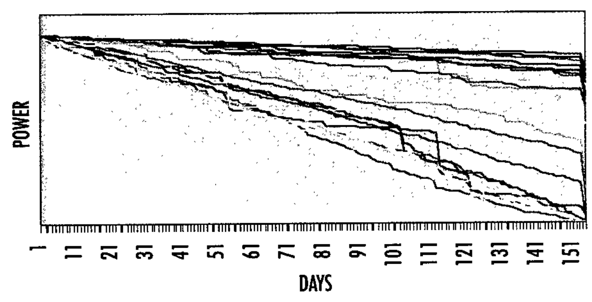


FIGURE 11. Vector-optimized field life.

Observations

The following observations are made regarding the simulation results:

1. The vector-optimized solution more than doubled field life as compared to the fixed-field solution.
2. The 2-of-2 detector has a longer life than the 1-2 detector. This is because the 2-of-2 detector has stringent initial detection rules, which translates to fewer reports and less communication as shown in Table 1.

3. Field life increased when the master node was moved from the edge of the field to the center of the field for the second field laydown. This is because when the master node is in the center of the field, there are more direct routes to the master node, which spreads out battery drain.
4. The vector-optimized solution has a longer field life than the single optimized solution. This is because changing the routes every 24 hours allows the battery drain to be spread more evenly across the field. With the vector-optimized solution, approximately one-third of the field will die on the same day.

CONCLUSIONS

In this paper, we have applied a stochastic optimization technique to adapt the thresholds of an autonomous sensor field and plan the communication routes. This stochastic optimization algorithm is known as evolutionary programming. The evolutionary program adapted the thresholds of a 2-of-2 detector for a set of sensors as well as a 1-2 detector. The algorithm is an evolutionary computation technique where an analytic solution is not attainable mathematically. Each sensor node in the 2-of-2 detector contained two thresholds to adapt, yielding four total thresholds to compute. The four thresholds are combined to meet a field-level probability of detection constraint and extend the life of a field of sensor nodes. Results show the benefits of adaptive threshold control in an autonomous sensor field by reducing communication costs and extending the life of the field by two.

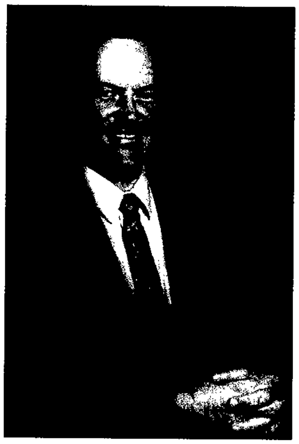
AUTHORS

Dale M. Klamer

MS in Mathematics, San Diego State University, 1972.
Current Research: Adaptive data fusion; advanced tracking; information assurance.

Barbara Dean

BS in Electrical Engineering, Widener University, 1989
Current Research: Software engineering for Navy sonar systems.



Mark W. Owen

MS in Electrical Engineering, California State University Long Beach, 1997
Current Research: Data fusion; signal processing; autonomous control.

REFERENCES

1. Hatch, M., M. Owen, et al. 1998. "Data Fusion Methodologies in the Deployable Autonomous Distributed System (DADS) Project," *Proceedings of the International Conference on Multisource-Multisensor Information Fusion*, (July), Las Vegas, NV, pp. 470-477.
2. Jahn, E., J. Kaina, and M. Hatch. 1999. "Fusion of Multi-Sensor Information from an Autonomous Undersea Distributed Field of Sensors," *Proceedings of the Second International Conference on Multisource-Multisensor Information Fusion*, (July), Sunnyvale, CA, pp. 4-11.
3. Shea, P. and M. Owen. 1999. "Fuzzy Control in the Deployable Autonomous Distributed System," *Proceedings of SPIE: Signal Processing, Sensor Fusion, and Target Recognition VIII 1999*, (Ivan Kadar, ed.), vol. 3720, (April), Orlando, FL.

4. Helstrom, C. W. 1968. *Statistical Theory of Signal Detection*, Pergamon Press, New York.

5. Klamer, D. and M. Owen. 2000. "Adaptive Threshold Control in an Autonomous Sensor Field," *Proceedings of SPIE: Signal and Data Processing of Small Targets 2000*, (Oliver Drummond, ed.), vol. 4048, (April), Orlando, FL.

6. Fogel, D. B. 1995. *Evolutionary Computation*, IEEE Press, New York.

❖

Use of One-Point Coverage Representations, Product Space Conditional Event Algebra, and Second-Order Probability Theory for Constructing and Using Probability-Compatible Inference Rules in Data-Fusion Problems

I. R. Goodman
SSC San Diego

ABSTRACT

This paper covers issues relating to the establishment of a sound and conditional probability-compatible rationale for generating linguistic-based inference rules concerning a population. By extending previous preliminary results, we detail, in a fully rigorous manner and within the confines of traditional probability theory, that a comprehensive technique can be derived that converts linguistic-based conditional information, couched only in fuzzy-logic terms, into naturally corresponding conditional probabilities. In turn, we demonstrate how such typically underconstrained conditional probabilities can be combined for suitable conclusions and decision-making, via a new use of second-order probability logic. This research is part of the ongoing SSC San Diego In-house Laboratory Independent Research FY 01 project CRANOF (a Complexity-Reducing Algorithm for Near-Optimal Fusion).

INTRODUCTION

Programmatics

This paper documents one aspect of the ongoing FY 01 In-house Laboratory Independent Research Project CRANOF (a Complexity-Reducing Algorithm for Near-Optimal Fusion), Project ZU014, with Principal Investigator, Dr. D. Bamber, and co-investigator, Dr. I. R. Goodman (both SSC San Diego), and with associate support from Dr. W. C. Torrez (SSC San Diego) and Prof. H. T. Nguyen (Department of Mathematical Sciences, New Mexico State University and U.S. Navy American Society for Engineering Education Fellow during summers at SSC San Diego). A preliminary version of this paper can be found in [1, section 3.3].

Background on Underconstrained Conditional Probability Problems

Philosophy of Approach and General Motivations

To improve the timeliness and accuracy of decision-supported human decision-making, one is faced with an array of crucial problems, including how to handle large amounts of incoming and uncertain information from disparate sources. These sources can be human-based or mechanical-based, and the information can arrive in different forms, such as qualitative and linguistic, numerical and statistical-probabilistic, or some mixture of both. At SSC San Diego, the CRANOF project addresses such crucial issues solely within the realm of statistics and probability. The issue of underconstrained or underspecified probabilities is treated by a novel use of second-order probabilities (i.e., probabilities of probabilities) in Bayesian framework. Underconstrained probabilities arise in a wide variety of problems, including quantitatively formulated rule-based systems, tracking and correlation, assessment of network intrusions, information retrieval, and simulation of human behavior in war games. This paper serves as a beginning extension of the capabilities of CRANOF to include linguistic-based information.

Quantitatively Formulated Rule-Based Systems

Consider quantitatively formulated rule-based systems, with the rules or conditional relations symbolized typically as $(a_1 \mid b_1)$, $(a_2 \mid b_2)$,...—read "if b_1, then a_1" (or equivalently, "a_1, given b_1," etc.), "if b_2, then a_2,"..., where events or sets a_1, b_1, a_2, b_2,... may themselves represent quite complicated logical combinations of simpler events or sets, and where it may or may not be known what logical relations exist among such events. Each such rule is also assigned quantitative reliability in the form of naturally corresponding conditional probabilities. Thus, for some otherwise unspecified probability measure P, rule $(a|b)$ is assigned value $P(a|b) = P(ab)/P(b)$, the conditional probability of a given b, using standard Boolean and probability notation and assuming antecedent probability $P(b) > 0$. Because typical rule $(a|b)$ is not perfect, in general $P(a|b) < 1$, but, on the other hand, one would expect $P(a|b)$ to be reasonably high. A common problem that such rule-based systems address is: Consider incoming information in the form of events, $d_1,\ldots, d_n$, possibly gleaned from different sources, such as d_1 = "visibility is up to 1 mile," d_2 = "winds between 15 mph and 30 mph," d_3 = "enemy movement detected last night in Sector C,"..., d_n = "political situation with enemy country Q at level R," and a collection of reasonably related rules, such as $(a_1|b_1)$, $(a_2|b_2)$,..., $(a_m|b_m)$, where the a_j, b_j involve not only parts or all of the d_j (or various logical combinations of them), but possibly other related events (or logical combinations of such). Then, one wishes to test for viability of possible decisions, based upon this information, such as c_1 = "fully successful attack by us can be accomplished by attacking in Sectors C or D," c_2 = "partially successful attack by us can be accomplished by attacking Sectors D or H,"... . Symbolically, one is considering the validity or degree of validity of the *entailment schemes* $G_i = [(a_1|b_1),\ldots, (a_n|b_n); (c_i|d)]$, $i = 1, 2,\ldots$, where $d = d_1\&\ldots\&d_n$ (conjunction of all data), and where $((a_1|b_1),\ldots, (a_n|b_n))$ can be considered the *premise set* of G_i and $(c_i|d)$ its *potential conclusion*. Ideally, one would like to know just what each $P(c_i|d)$ would be, based on having either, say, the *exact threshold situation* holding, i.e., $P(a_j|b_j) = t_j$, $j = 1,\ldots, n$, or, the *lower bound threshold* situation holding, i.e., having $P(a_j|b_j) \geq t_j$, where all the thresholds t_j are known or estimable in either situation. However, in general, it is readily demonstrated that the n equalities (or inequalities) are not enough to determine P and/or $P(c_i|d)$ completely. Thus, one is faced with the problem of best estimating, in some sense, just what P and/or $P(c_i|d)$ should be.

Adams' Approach to Analyzing Quantitatively Formulated Rule-Based Systems

In a series of papers [2, 3], E. W. Adams proposed, in effect, the estimate of $P(c_j|d)$ to be a pessimistic one in the form of his "minimum conclusion" function, using multivariable abbreviation t_J for $(t_j)_{j \text{ in } J}$, $(a|b)_J$ for $(a_j|b_j)_{j \text{ in } J}$, $P(a|b)_J \geq t_J$ for $P(a_j|b_j) \geq t_j$, j in J, 1_J for column vector of all 1's indexed by J, etc.,

$$\text{estimate}_{HPL} \text{ of } (P(c_i|d) \text{ from } G_i) = \text{minconc}(G_i)(t_J) = \inf\{P(c_i|d)\text{: for all possible probability measures P such that } P(a|b)_J \geq t_J\}, \quad (1)$$

with $P(c_i|d)$ for the exact threshold situation analogously estimated. The subscript $(\ )_{HPL}$ is used to indicate "High Probability Logic," since Adams also introduced the idea of an entailment scheme being *HP-valid* or *HP-invalid*, which, in the case of any G_i here simply means for the former that

$$G_i \text{ is HPL-valid} \quad \text{iff} \quad \underset{(t_J \uparrow 1_J)}{\text{limit}}(\text{minconc}(G_i)(t_J)) = 1. \tag{2}$$

But, unfortunately, both the minconc function and its limiting forms to test for HPL-validity/invalidity produce a number of results very much at odds with commonsense reasoning, including the fact that three very fundamental entailment schemes, *transitivity* (or *hypothetical syllogism*) [(a|b), (b|c); (a|c)] (the heart of any rule-based system); *contraposition* [(a|b); (b′|a′)]; and *strengthening of antecedent* [(a|b); (a|bc)] are all HPL-invalid. In fact, one can find P's that satisfy their premise thresholds for any choice of t_J close to (but not exactly equal to) 1_J, but for which the corresponding conclusion probabilities are arbitrarily close to (or actually equal to) 0. Moreover, more generally, Eq. (2) can be complemented by the fact that any

$$G_i \text{ is HPL-invalid} \quad \text{iff} \quad \underset{(t_J \uparrow 1_J)}{\text{limit}}(\text{minconc}(G_i)(t_J)) = 0. \tag{3}$$

Finally, Adams pointed out another type of validity, CPL (Certainty Probability Logic), that, although still based on the minconc function, can be characterized as "too optimistic" in contrast with HPL, whereby the criterion is

$$G_i \text{ is CPL-valid} \quad \text{iff} \quad \text{minconc}(G_i)(1_J)) = 1. \tag{4}$$

Close connections exist between CPL validity/invalidity (the latter satisfying a relation analogous to that of Eq. (3)) and that of CL (classical logic) validity or invalidity, noting

$$G_i \text{ is CL-valid} \quad \text{iff} \quad \&(b' \vee ab)_J \leqslant d' \vee c_i d. \tag{5}$$

(For further analysis, criticism, and extension of Adams' ideas, see [3].)

CRANOF Approach to Analyzing Quantitative Rule-Based Systems and Other Underconstrained Probability Problems

The previous conclusions show that the minconc function is not a reasonable measure (for reasonably high thresholds) of the degree of validity/invalidity of an entailment scheme and also show that the HP-validity/invalidity test is too stringent. Therefore, it seemed natural to replace the extremal minconc function by the more moderating *meanconc* function (well-justified from decision analysis in the form of conditional expectation and justified as always admissible, least-squares error, etc.—see any standard texts such as Rao [4] or Wilks [5]) within a Bayesian framework, where the unknown probability measure P here is treated as a random quantity with some appropriately assigned prior distribution, subject to the given premise set threshold constraints. Utilizing additional new theoretical results [6], an "optimal" choice of prior or priors essentially must come from the well-known Dirichlet family of distributions. It should be noted that, unlike the minconc function, the meanconc function in the unity-limiting threshold case can take on nontrivial values and, in a natural sense, at any fixed threshold level, provides a reasonable measure of degree of validity of that entailment scheme under consideration. In particular, in full agreement with commonsense reasoning, transitivity, contraposition, and strengthening of antecedent are all SOPL-valid, where SOPL stands for Second-Order Probability Logic and where one defines validity of any G_j as

$$G_i \text{ is SOPL-valid} \quad \text{iff} \quad \underset{(t_J \uparrow 1_J)}{\text{limit}}(\text{meanconc}(G_i)(t_J)) = 1, \qquad (6)$$

SOPL-validity depending on some degree, of course, on the particular choice of prior for P. However, it has been pointed out (Bamber [7] and personal communications) that the limit in Eq. (4) remains the same as if the prior of P is a uniform distributional one, when the corresponding probability density function is bounded uniformly above and below (from zero) over its *natural* domain (again, see references).

Also, see [8] for additional background on both the theoretical structure of the meanconc function and its practical implementational form CRANOF—whereby a significant reduction in the complexity of computing meanconc$(G_i)(t_J)$ is achieved by, in effect, reducing the premise set of G_i to a single constraint, also taking into account the unity-limiting threshold behavior of meanconc ([7]). Finally, Table 1 is presented below to illustrate a few typical evaluations of meanconc(G) for relatively simple entailment schemes G with P assigned a uniform prior distribution [8].

TABLE 1. Abridged table of calculations of degree-of-entailment functions, minconc and meanconc, for fixed threshold levels, and a comparison of CPL-, SOPL-, and HPL-validities for different types of entailment schemes.

Name of Entailment Scheme $D = [(a\mid b)_J; (c\mid d)]$	Given Levels of Premises: $P(a\mid b)_J = t_J$, for otherwise arbitrary prob. meas. P	minconc$(D)(t_J)$ (inequality threshold form)	meanconc$(D)(t_J)$, assuming uniform prior for P's (exact threshold form)	D is CPL-valid?	D is SOPL-valid?	D is HPL-valid?
Cautious Monotonicity: $[(a\mid b),(c\mid b); (a\mid bc)]$	$P(a\mid b) = s$, $P(c\mid b) = t$	$\geq \max(s+t-1,0)$	$\geq \max(s+t-1,0)$	YES	YES	YES
Transitivity: $[(a\mid b), (b\mid c); (a\mid c)]$	$P(a\mid b) = s$, $P(b\mid c) = t$	0	$= st + (1-t)/2 - p(s,t)/q(s,t)$, $p(s,t) = s(1-s)(2s-1)t(1-t^2)$, $q(s,t) = t+2t^2+ (s(1-s)(1-t)(2+3t-t^2)$	YES	YES	NO
Contraposition: $[(a\mid b); (b'\mid a')]$	$P(a\mid b) = t$	0	$1/t + \dfrac{(1-t)\log(1-t)}{t^2}$	YES	YES	NO
Positive Conjunction: $[(a\mid b),(a\mid c); (a\mid bc)]$	$P(a\mid b) = t$, $P(a\mid c) = t$	0	$(1+t)/3 + [((1+t)(2-t)/(3t))\,\theta(t)]$, $\theta(t)$ $= (t^2/4)[\log((2-t)/t)]/(1-t)$ $- ((1-t)^2/4)\cdot\log((1+t)/(1-t))$	YES	YES	NO
Nixon Diamond: $[(ab\mid c),(d\mid a),(d'\mid b); (d\mid c)]$	$P(ab\mid c) = s$, $P(d\mid a) = t$, $P(d'\mid b) = t$	0	1/2	YES	NO	NO
Abduction: $[(a\mid b), a; b]$	$P(a\mid b) = s$, $P(a) = t$	0	If $s \geq t : t/(2s)$, If $s < t : \dfrac{t^3\, s(1-t)^2}{2(t^2 - 2st + s)^2}$	NO	NO	NO

EXTENDING APPLICABILITY OF CRANOF TO LINGUISTIC-BASED SYSTEMS

In considering linguistic-based information in rule-based systems and in formulating the linguistic analogue of the underconstrained conditional (including unconditional) probability problem, the role of fuzzy logic comes immediately to mind. This is based in part on the great practical success of fuzzy logic in running systems such as elevators, washing machines, etc., and on the now very large body of scientific literature supporting the modeling of linguistic information, relations, and decision processes via fuzzy logic. (See, e.g., past *Proceedings of IEEE International Conferences on Fuzzy Systems* or the *Proceedings of the Joint Conference on Information Sciences*, as well as basic texts, such as Dubois & Prade's now classic treatise [9] and Nguyen & Walker's [10].)

On the other hand, there still exists a lively controversy considering the merits of using probability theory and techniques in place of fuzzy logic and vice versa. (See Goodman's summary and listing of literature papers directly involved in this controversy [11].) This leads to the following area in which this author and H. T. Nguyen have played some role over the past several years: *the issue of the possible direct connection between fuzzy logic and probability theory* [12, 13, and 1]. Until this is completely resolved, it is this author's opinion that a comprehensive view of data fusion, which both theoretically and practically integrates linguistic-based information with probabilistic-based information, will not be achieved. In particular, this applies to rule-based systems, where the fuzzy logic community has developed a common approach that is claimed to be more satisfactory than any probability approach.

This paper once again points out the existence of deep, but tractable, relations among fuzzy logic, linguistic-based principles, probability theory, and commonsense reasoning mainly through the use of two basic mathematical tools: SOPL/CRANOF (as briefly described in the first section), and the representation theory of fuzzy sets by the one-point coverages of random sets (see [12, 13]) in conjunction with other recently developed mathematical tools (*conditional and relational event algebra* [14; 15, section 3]). In particular, homomorphic-like relations were established, connecting fuzzy-logic concepts and corresponding random-set concepts, where each fuzzy-set membership function involved is, in effect, interpreted as the weakest way to specify any of a class of corresponding random subsets of the fuzzy set's domain. These relations include natural random-set interpretations of various combinations of fuzzy-logic operators and Zadeh's well-known "extension theorem." This time, these connections are extended to include the formulation and use of inference rules obtained from a population of interest. The results presented here extend preliminary efforts provided in Goodman & Nguyen [1], where it was demonstrated that one type of fuzzy-logic approach to the modeling of inference rules for a population, relative to a given collection of attributes, using the ratio of fuzzy cardinalities or averaged membership level of the attributes, could also be interpreted in a probability framework. In addition, by using similar techniques, it is shown how other fuzzy-logic concepts, commonly thought of as not directly relating to probability, may now also be put into a complete probabilistic setting, including the illustration for normalization of membership functions.

MATHEMATICAL RESULTS ESTABLISHING GENERAL FUZZY LOGIC POPULATION CONDITIONING PROBLEM AS AN UNDERCONSTRAINED CONDITIONAL PROBABILITY PROBLEM TREATABLE VIA SOPL/CRANOF

As in the previous sections, standard Boolean algebra and probability theory notation will be employed, with [0,1] indicating unit interval; {0,1} indicating the two element set containing 0, 1; **R** indicating the real (or Euclidean) line and $\mathbf{R}^m$ indicating the real (or Euclidean m-space), P(D) indicating the power class of D (sometimes written 2^D—the class of all subsets of D), etc. "Equal by definition" is denoted as $=_d$. For background on copulas, see Schweizer & Sklar [16] and the recent excellent monograph by Nelsen [17]. Recall that copulas are any joint cdf's (cumulative probability distribution functions), all of whose one-dimensional marginal cdf's correspond to identical uniformly distributed rv's (random variables) over [0,1].

Theorem 1. Modification of Goodman [18]

Let D be a finite set, f, g:D→[0,1] any two fuzzy-set membership functions, and cop: $[0,1]^{D\times D}\to[0,1]$ any copula with that domain, with (x,y)-marginal copulas indicated by, e.g., $\mathrm{cop}_{x,y}$, x, y in D, etc. Then:

(i) There is a probability space (**Ω**,B,P) and a joint collection of 0-1-valued rv's, $Z_{f,x}$, $Z_{g,y}$:**Ω**→{0,1}, for all x, y in D with overall joint cdf $F_{f,g,cop} = \mathrm{cop}_o((F_{f,x})_{x \text{ in } D}, (F_{g,y})_{y \text{ in } D})$: $\mathbf{R}^{D\times D}\to[0,1]$ (via Sklar's Theorem [16]), and, indicating the joint marginal (x,y)-components of cop, as $\mathrm{cop}_{x,y}$, the joint cdf of ($Z_{f,x}$, $Z_{g,y}$) is, correspondingly, $F_{f,g,cop,x,y}(., ..) = \mathrm{cop}_{x,y}{}^o(F_{f,x}(.), F_{g,y}(..))$, where ^{o} indicates functional composition and $F_{f,x}$, $F_{g,y}$ are each one-dimensional cdf's corresponding to mass-point probability functions $h_{f,x}$, $h_{g,y}$, respectively, where

$$P(Z_{f,x} = 1) = h_{f,x}(1) = f(x);\ P(Z_{f,x} = 0) = h_{f,x}(0) = 1\text{-}f(x);$$
$$P(Z_{g,y} = 1) = h_{g,y}(1) = g(y); P(Z_{g,y} = 0) = h_{g,y}(0) = 1\text{-}g(y); \quad (7)$$

whence

$$F_{f,x}(s)=\begin{cases} 0, \text{ if } s < 0, \\ 1\text{-}f(x), \text{ if } 0 \leqslant s < 1, \\ 1, \text{ if } 1 \leqslant s; \end{cases} \quad F_{g,y}(s)=\begin{cases} 0, \text{ if } s < 0, \\ 1\text{-}g(y), \text{ if } 0 \leqslant s <1, \\ 1, \text{ if } 1 \leqslant s; \end{cases} \text{ all x, y in D} \quad (8)$$

(ii) Define random sets S(f, cop), S(g, cop):**Ω**→P(D), S(f, g, cop): **Ω**→P(D) ×P(D) as follows, for each ω in **Ω**:

$$S(f, g, cop)(\omega) = S(f, cop)(\omega)\times S(g, cop)(\omega) = \{(x,y): x, y \text{ in } D, Z_{f,x}(\omega)\, Z_{g,y}(\omega) = 1\};$$
$$S(f, cop)(\omega) = \{x: x \text{ in } D, Z_{f,x}(\omega) = 1\};\quad S(g, cop)(\omega) = \{y: y \text{ in } D, Z_{g,y}(\omega) = 1\}; \quad (9)$$

whence, by straightforward combinatoric considerations, the entire probability distributions of the marginal random subsets of D, S(f, cop), S(g, cop), as well as the joint random subset of D×D, S(f, g, cop), are completely determined.

(iii) For any x, y in D, the following equality of one-point coverage events hold:

$$(x \text{ in } S(f, \text{cop})) = (Z_{f,x} = 1)\ ;\ (y \text{ in } S(g, \text{cop})) = (Z_{g,y} = 1); \qquad (10)$$

$$((x,y) \text{ in } S(f, g, \text{cop})) = (x \text{ in } S(f, \text{cop}))\ \&\ (y \text{ in } S(g, \text{cop})) = (Zf,x = 1)\ \&\ (Z_{g,y} = 1). \qquad (11)$$

(iv) For any x, y in D, the following *one-point coverage representations* for f, g hold:

$$P(x \text{ in } S(f, \text{cop})) = P(Z_{f,x} = 1) = f(x)\ ;\ P(y \text{ in } S(g, \text{cop})) = P(Z_{g,y} = 1) = g(y); \qquad (12)$$

$$\begin{aligned} P((x \text{ in } S(f, \text{cop}))\ \&\ (y \text{ in } S(g, \text{cop}))) &= P((Z_{f,x} = 1)\ \&\ (Z_{g,y} = 1)) \\ &= 1 - P(Z_{f,x} = 0) - P(Z_{g,y} = 0) + P((Z_{f,x} = 0)\ \&\ (Z_{g,y} = 0)) \\ &= 1 - P(Z_{f,x} = 0) - P(Z_{g,y} = 0) + P((Z_{f,x} \leq 0)\ \&\ (Z_{g,y} \leq 0)) \\ &= 1 - (1-f(x)) - (1-g(y)) + F_{f,g,\text{cop}_{x,y}}(0, 0) \\ &= f(x) + g(y) - 1 + \text{cop}_{x,y}(1-f(x), 1-g(y)) \\ &= f(x) + g(y) - \text{cocop}_{x,y}(f(x), g(y)) \\ &=_d \text{cop}_{x,y}{}^\wedge(f(x), g(y)), \end{aligned} \qquad (13)$$

where we use the relation

$$\begin{aligned} F_{f,g,\text{cop}_{x,y}}(0, 0) &= \text{cop}_{x,y}{}^\circ(F_{f,x}(0), F_{g,y}(0)) \\ &= \text{cop}_{x,y}{}^\circ(h_{f,x}(0), h_{g,y}(0)) \\ &= \text{cop}_{x,y}{}^\circ(1-f(x), 1-g(y)) \end{aligned}$$

and where the functions cocop, cop^ are called the *cocopula, survival copula*, respectively, of cop (the latter apparently being the special designation of Nelsen for modular transform [17, section 2.6]), where, for any s, t in [0,1]:

$$\text{cocop}(s, t) =_d 1 - \text{cop}(1-s, 1-t)\ ;\ \text{cop}^\wedge(s, t) =_d s+t - \text{cocop}(s,t). \qquad (14)$$

(v) Specializing (iv) for x = y in D arbitrary,

$$P(x \text{ in } S(f, \text{cop}) \cap S(g, \text{cop})) = P((x,x) \text{ in } S(f, g, \text{cop})) = \text{cop}_{x,y}{}^\wedge(f(x), g(x)). \qquad (15)$$

(vi) As copula cop is allowed to vary arbitrarily, the full solution set of distribution-distinct random subsets of D that are one-point coverage equivalent to f, g, respectively in the sense of Eq. (12), is exhausted. ■

Remark 1. Note first that cocop is the DeMorgan transform of cop—so that if one thinks of cop as a generalized conjunction or "and" operator—as in fuzzy logic (with the usual desirable properties of being nondecreasing in its arguments and having appropriate boundary properties when one of the arguments is 0 or 1), then, naturally, cocop can be thought of as a general disjunction or "or" operator. Nelsen [17, section 2], shows that the survival copula is always a legitimate copula and shows the characterization

$$\text{cop}^\wedge = \text{cop} \quad \text{iff} \quad \text{cop is } \textit{radially} \text{ symmetric}, \qquad (16)$$

where the latter means that the joint r.v. Y represented by cop is such that $Y - (1/2, 1/2)$ and $(1/2, 1/2) - Y$ have the same distribution. In particular, radial symmetry—and hence the validity of Eq. (16)—holds for all Gaussian copulas $\Psi_\rho\,(\Psi\text{-}1(.), \Psi\text{-}1(..))$, where Ψ_ρ is the joint cdf of distribution Gaussian $(0_2, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$ and Ψ is the cdf of the standardized one-dimensional Gaussian distribution Gaussian (0,1) and all of Frank's Archimedean copula family [17], [19] (i.e., associative, commutative with cop(s,s) < s, for 0 < s < 1)—which includes the copulas prod and minsum, as well as the special copula min, where for all s, t in [0,1], min, prod are the usual arithmetic minimum and product of s, t, respectively, while minsum(s,t) is given as

$$\mathrm{minsum}(s,t) = \min(s+t-1,\ 0). \tag{17}$$

■

Theorem 2. Extension of Goodman & Nguyen [13]

Suppose that D is a finite set, f, g:D→[0,1] are any two fuzzy set membership functions, cop: $[0,1]^{D\times D}\to[0,1]$ is any copula with that domain, and w:D→[0,1] is a probability function. Define

$$((f|g)_{\mathrm{cop},w} =_d \sum_{\mathrm{xinD}} (\ w(x)\cdot \mathrm{cop}^\wedge(f(x),\ g(x)))\ \Big/ \sum_{\mathrm{xinD}} (\ w(x)\cdot g(x)). \tag{18}$$

Then, in the sense of Theorem 1, there is a probability space (Ω,B,P) and random sets S(f, cop), S(g, cop): Ω →*P*(D), S(f,g, cop): with the one-point coverage relations holding as in Eqs. (12), and, without loss of generality, there exists a random variable V:Ω→D, independent of S(f, g, cop), and hence of S(f, cop), S(g, cop), such that the probability function of V is w, so that

$$(f|g)_{\mathrm{cop},w} = P(a_{f,\ \mathrm{cop}}\ |\ b_{g,\ \mathrm{cop}}), \tag{19}$$

an ordinary conditional probability, where events $a_{f,cop}$, $b_{g,cop}$ in B are defined as the two-stage randomization events

$$a_{f,\ \mathrm{cop}} =_d\ (V \text{ in } S(f,\ \mathrm{cop}))\ ,\quad b_{g,\ \mathrm{cop}} = (V \text{ in } S(g\ \mathrm{cop}))\ , \tag{20}$$

so that in reduced form,

$$P(a_{f,\ \mathrm{cop}}\ |\ b_{g,\ \mathrm{cop}}) = P(a_{f,\ \mathrm{cop}} \,\&\, b_{g,\ \mathrm{cop}}\ |\ b_{g,\ \mathrm{cop}}) = P(V \text{ in } S(f,\ \mathrm{cop}) \cap S(g,\ \mathrm{cop}))\ /\ P(V \text{ in } S(g\ \mathrm{cop})). \tag{21}$$

Proof: Use the usual conditioning property of probabilities, independence of V, and Eq. (11) at each outcome of r.v. V,

$$\begin{aligned} P(V \text{ in } S(f,\mathrm{cop}) \text{ and } V \text{ in } S(g,\ \mathrm{cop}))\ &= E_V(P(V \text{ in } S(f,\mathrm{cop}) \text{ and } V \text{ in } S(g,\ \mathrm{cop})\ |\ V)) \\ &= E_V(\mathrm{cop}^\wedge(f(V),\ g(V))) = \sum_{\mathrm{xinD}} (\ w(x)\cdot \mathrm{cop}^\wedge(f(x),\ g(x))). \end{aligned} \tag{22}$$

Similarly (and more simply), now using Eq. (12) in place of Eq. (13),

$$P(V \text{ in } S(g,\ \mathrm{cop})) = E_v(P(V \text{ in } S(g,\mathrm{cop})\ |\ V)) = E_v(g(V)) = \sum_{\mathrm{xinD}} (\ w(x)\cdot g(x)). \tag{23}$$

The desired results hold by dividing Eq. (22) by Eq. (23). ■

Remark 2 and an Example. In Theorem 2, for the special case of w corresponding to a uniform distribution over *population* D, canceling the 1/card(D) factor, and usually—but not always choosing cop to be either min or prod—the numerator of the quantity $(f|g)_{\mathrm{cop},\ w}$ reduces to the popular fuzzy-logic concept of the *fuzzy cardinality* of f "and" g for population D, i.e., to what extent the entire population D has characteristics described by f "and" g, while, similarly, the denominator represents the fuzzy cardinality of g (by itself) for population D. In turn, the arithmetic division of these, i.e., the quantity $(f|g)_{cop,w}$ becomes the *relative fuzzy cardinality* of f "and" g for D *compared to fuzzy cardinality* of g for D, i.e., the overall fuzzy conditioning of f to g with respect to population D. The latter, beginning with Zadeh's ideas [20, 21], followed by Dubois & Prade's modifications [22], and Kosko's related concept of *fuzzy subsethood* [23], are used ubiquitously in the fuzzy-logic community for reasoning. In this process, one considers the premise set of a particular linguistic entailment of interest, the latter being formally the same as the probability-framed previous $G_i = [(a|b)_J; (c_i|d)]$, but now where each $(a_j|b_j)$ is replaced by a fuzzy conditional—in its general form *the same* as $(f_j|g_j)_{cop,w}$ —formed as in Eq. (18), now with f replaced by f_j, g by g_j (for possibly pre-logically compounded fuzzy-set membership functions), j in J; and with similar remarks applicable to the potential conclusion $(c_i\,|d)$ replaced by $(f_{o,i}|g_o)_{cop,w}$, for some fuzzy sets $f_{o,i}$, g_o, etc. But, Theorem 2 (with suitable modifications, where required) essentially shows that any such $(f_j|g_j)_{cop,w} = P(a_{fj,\,cop} \mid b_{gj,cop})$, with a similar relation holding the potential conclusion. Moreover, the variability of P subject to whatever arbitrary but fixed levels t_J are set for the premise collection holds in the same meaningful manner as in the case where one began the problem in a probability framework, i.e., for typical entailment schemes of the form G_i. As an application of this, suppose one considers the transitivity scheme, which Zadeh has also considered and modeled his premise set as indicated above, but has used a method solely developed within fuzzy logic for determining what the appropriate conclusion should be [21]. Thus, three attributes are present, where, e.g., population D here is the set of all enemy ships in area A, "ships with type 1 weapons onboard" corresponds to known or estimated fuzzy-set membership function f over D; "ships with elongated hulls" corresponds to known or estimated fuzzy-set membership function g over D; "ships with signature pattern Q" corresponding to known or estimated fuzzy-set membership function h over D. Moreover, other truth modifiers may be present, such as "it is mostly true," "it is somewhat true," etc. Here, for simplicity, suppose for the premise set, one actually has "it is highly true that the enemy ships in A with signature pattern Q have elongated hulls," "it is moderately likely that an enemy ship in A with an elongated hull has type 1 weapons onboard." Can one conclude "it is x-likely that an enemy ship in A with signature pattern Q has type 1 weapons onboard," where the degree of truth x is to be determined? Assume that "it is highly true" is represented by a known or estimated fuzzy-set membership function M over [0,1], which is monotone increasing, "it is moderately likely" is also represented by a (different—not as steep toward 1 as M, etc.) known or estimated fuzzy-set membership function N over [0,1], where M(r) = N(r) = r, for r = 0 or 1. Hence, for any arbitrary levels s, t in [0,1], the conditional fuzzy relations here are, for some choice of copula and population weighting function w,

$$\begin{aligned} M((f|g)_{cop,w}) = s\,,\ N((g|h)_{cop,w}) = t \quad &\text{iff} \quad (f|g)_{cop,w} = M^{-1}(s)\,,\ (g|h)_{cop,w} = N^{-1}(t) \\ &\text{iff, using Theorem 2,}\ \ P(a_{f,\,cop} \mid b_{g,cop}) = M^{-1}(s), \\ &\quad P(b_{g,\,cop} \mid c_{h,cop}) = N^{-1}(t). \end{aligned} \tag{24}$$

Thus, for any given levels s, t, one can now consider the SOPL-estimate of the potential conclusion for transitivity, $P(a_{f,\,cop} \mid b_{g,cop})$, with respect to the premise set above at thresholds s, t, where the entire entailment scheme is

$$G = [(a_{f,\,cop} \mid b_{g,cop}), (b_{g,\,cop} \mid c_{h,cop}); (a_{f,\,cop} \mid c_{h,cop})]; \tag{25}$$

$$\begin{aligned}&\mathrm{meanconc}(G)(M^{-1}(s), N^{-1}(t)) = E_P(P(a_{f,\,cop} \mid c_{h,cop}) \mid \\ &P(a_{f,\,cop} \mid b_{g,cop}) = M^{-1}(s)\ , P(b_{g,\,cop} \mid c_{h,cop}) = N^{-1}(t)).\end{aligned} \tag{26}$$

In turn, Table 1 shows that under a uniform distributional assumption on what P could be, subject to its constraints in the premise set of G, for any given s, t in $[1/2, 1]$

$$\mathrm{meanconc}(G)(M^{-1}(s), N^{-1}(t)) = \rho(M^{-1}(s), N^{-1}(t)),$$

where, for any s, t in $[1/2, 1]$,

$$\begin{aligned}&\rho(s,t) =_d st + (1-t)/2 - p(s,t)/q(s,t);\ p(s,t) =_d s(1-s)(2s-1)t(1-t^2); \\ &q(s,t) =_d t+2t^2 + (s(1-s)(1-t)(2+3t-t^2)),\end{aligned} \tag{27}$$

where,

$$\rho(s,t) \approx \rho_o(s,t) =_d st + (1-t)/2\ , \text{ for values of s, t sufficiently close to 1.} \tag{28}$$

Hence, the posterior conditional (given the premise constraints for any s, t) is approximately equal to $\rho_o(M^{-1}(s), N^{-1}(t))$, which can be interpreted also as a truth modifier with respect to two variables, noting its limit is unity as s, t approach unity, etc. Of course, all of the above applies to any fuzzy-logic entailment scheme relative to the original premise sets utilizing overall fuzzy conditioning for some population D.

Remark 3. In the same spirit of Theorem 2, other fuzzy-logic concepts can now be fully interpreted. Due to space limitations, only the example of fuzzy normalization will be considered here. In this situation, a fuzzy membership function, say, $f{:}D \to [0,1]$ is given, followed by its normalization function $\mathrm{norm}(f){:}D \to [0,1]$, which is now obviously a legitimate probability function over finite population D, where

$$\mathrm{norm}(f) = \Big(1/\sum_{x in D}(f(x))\ \Big)\cdot f\ . \tag{29}$$

But, if one considers, à la Theorem 1, for any choice of copula cop, a probability space $(\Omega,\mathcal{B},P)$, for which, without loss of generality, there is both a random set $S(f, cop){:}\Omega \to \mathcal{P}(D)$ and an independent random variable $V{:}\Omega \to D$ uniformly distributed over D, with the one-point coverage relation holding

$$P(x \text{ in } S(f, cop)) = f(x), \text{ all } x \text{ in } D, \tag{30}$$

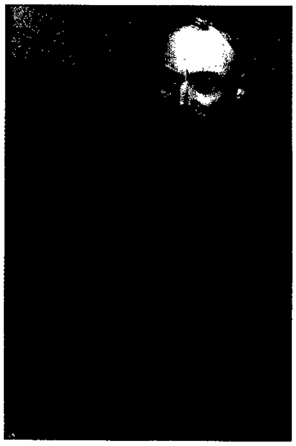
for any x in D, specializing Eq. (23) with g replaced by f,

$$\begin{aligned}P(V = x \mid V \text{ in } S(f, cop)) &= P(V = x \text{ and } x \text{ in } S(f, cop)) \,/\, P(V \text{ in } S(f, cop)) \\ &= ((1/\mathrm{card}(D))\cdot f(x)) \,/ \sum_{x in D} (1/\mathrm{card}(D))\cdot g(x)\Big) = \ \mathrm{norm}(f)(x),\end{aligned} \tag{31}$$

showing fuzzy normalization is actually a simple conditional probability restriction of the two-stage randomization for one-point coverages. A future paper will deal with related issues.

REFERENCES

1. Goodman, I. R. and H. T. Nguyen. 1999. "Application of Conditional and Relational Event Algebra to the Defining of Fuzzy Logic Concepts," *Proceedings of Signal Processing, Sensor Fusion & Target Recognition VIII*, Society of Photo-Optical Instrumentation Engineers (SPIE), vol. 3720, pp. 37–46.
2. Adams, E. W. 1986. "On the Logic of High Probability," *Journal of Philosophical Logic*, vol. 15, pp. 255–279.
3. Adams, E. W. 1996. "Four Probability-Preserving Properties of Inferences," *Journal of Philosophical Logic*, vol. 25, pp. 1–24.
4. Rao, C. R. 1973. *Linear Statistical Inference & Its Applications, 2nd Ed.*, Wiley, New York, NY.
5. Wilks, S. S. 1963. *Mathematical Statistics*, Wiley, New York, NY.
6. Goodman, I. R. and H. T. Nguyen 1999. "Probability Updating Using Second-Order Probabilities and Conditional Event Algebra," *Information Sciences*, vol. 121, pp. 295–347.
7. Bamber, D. 2000. "Entailment with Near Surety of Scaled Assertions of High Conditional Probability," *Journal of Philosophical Logic*, vol. 29, pp. 1–74.
8. Bamber, D. and I. R. Goodman. 2000. "New Uses of Second-Order Probability Techniques in Estimating Critical Probabilities in Command and Control Decision-Making," *Proceedings of the 2000 Command & Control Research & Technology Symposium*, Naval Postgraduate School, http://www.dodccrp.org/2000CCRTS/cd/html/pdf_papers/Track_4/124.pdf.
9. Dubois, D. and H. Prade. 1980. *Fuzzy Sets & Systems: Theory and Applications*, Academic Press, New York, NY.
10. Nguyen, H. T. and E. A. Walker. 1997. *A First Course in Fuzzy Logic*, CRC Press, New York, NY.
11. Goodman, I. R. 1998. "Random Sets and Fuzzy Sets: a Special Connection," *Proceedings of the International Conference on Multisource-Multisensor Information Fusion (Fusion'98)*, vol. 1, pp. 93–100.
12. Goodman, I. R. and H. T. Nguyen. 1985. *Uncertainty Models for Knowledge-Based Systems*, North-Holland Press, Amsterdam.
13. Goodman, I. R. and G. F. Kramer. 1997. "Extension of Relational and Conditional Event Algebra to Random Sets with Applications to Data Fusion," in *Random Sets: Theory & Applications* (J. Goutsias, R. P. Mahler, and H. T. Nguyen, eds.), Springer, New York, NY, pp. 209–242.
14. Goodman, I. R. and H. T. Nguyen. 1995. "Mathematical Foundations of Conditionals and Their Probabilistic Assignments," *International Journal of Uncertainty, Fuzziness & Knowledge-Based Systems*, vol. 3, no. 3 (September), pp. 247–339.
15. Goodman, I. R., R. P. Mahler, and H. T. Nguyen. 1997. *Mathematics of Data Fusion*, Kluwer Academic, Dordrecht, Holland.
16. Schweizer, B. and A. Sklar. 1983. *Probabilistic Metric Spaces*, North-Holland, Amsterdam.
17. Nelsen, R. B. 1999. *An Introduction to Copulas* (Lecture Notes in Statistics, no. 139), Springer, New York, NY.



I. R. Goodman

Ph.D. in Mathematics, Temple University, 1972

Current Research: Mathematical foundations of data fusion via conditional probabilistic logic; Boolean conditional event algebra; one-point random set representations of fuzzy logic.

18. Goodman, I. R. 1994. "A New Characterization of Fuzzy Logic Operators Producing Homomorphic-Like Relations with One Point Coverage of Random Sets," in *Advances in Fuzzy Theory & Technology*, (P. P. Wang, ed.), Duke University, Durham, NC, vol. 2, pp. 133–159.

19. Frank, M. J. 1979. "On the Simultaneous Associativity of F(x,y) and x+y-F(x,y)," *Aequationes Mathematicae*, vol. 19, pp. 194–226.

20. Zadeh, L. A.1985. "Syllogistic Reasoning as a Basis for Combination of Evidence in Expert Systems," *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-85)*, vol. 1, pp. 417–419.

21. Zadeh, L. A. 1978. "PRUF: a Meaning Representation Language for Natural Languages," *International Journal of Man–Machine Studies*, vol. 10, pp. 395–460.

22. Dubois, D. and H. Prade. 1988. "On Fuzzy Syllogisms," *Computational Intelligence*, vol. 4, pp. 171–179.

23. Kosko, B. 1992. *Neural Networks and Fuzzy Systems*, Prentice-Hall, Englewood Cliffs, NJ.

❖

On Knowledge Amplification by Structured Expert Randomization (KASER)

Stuart H. Rubin
SSC San Diego

ABSTRACT

We define Knowledge Amplification by Structured Expert Randomization (KASER). A KASER can automatically acquire a virtual rule space exponentially larger than the actual rule space and with an exponentially decreasing nonzero likelihood of error. The KASER cracks the knowledge acquisition bottleneck in intelligent systems by amplifying user-supplied knowledge. This enables the construction of an intelligent system, which is creative, fail-soft, learns over a network, and otherwise has enormous potential for automated decision-making.

INTRODUCTION TO RANDOMIZATION

The theory of randomization was first published by Chaitin and Kolmogorov [1] in 1975. Their work may be seen as a consequence of Gödel's Incompleteness Theorem [2] in that it shows were it not for essential incompleteness, a universal knowledge base could, in principle, be constructed—one that need employ no search other than referential search. Lin and Vitter [3] proved that learning must be domain-specific to be tractable. The fundamental need for domain-specific knowledge is in keeping with Rubin's proof of the Unsolvability of the Randomization Problem [4]. This paper went on to introduce the concept of knowledge amplification. Production rules are expressed in the form of situation action pairs. Such rules, once discovered to be in error, are corrected through acquisition. Conventionally, a new rule must be acquired for each correction. This is linear learning.

The acknowledged key to breakthroughs in the creation of intelligent software is cracking the knowledge acquisition bottleneck [5]. Learning how to learn is fundamentally dependent on representing the knowledge in the form of a society of experts. Minsky's seminal work here led to the development of intelligent agent architectures [6]. Furthermore, Minsky [7] and Rubin [4] independently provided compelling evidence that the representational formalism itself must be included in the definition of domain-specific learning if it is to be scalable.

A KASER is defined to be a knowledge amplifier that is based on the principle of structured expert randomization. A Type I KASER is one where the user supplies declarative knowledge in the form of a semantic tree using single inheritance.

A Type II KASER can automatically induce this tree through the use of randomization and set operations on property lists, which are acquired by way of database query and user-interaction. An overview of a Type II KASER is provided below. Unlike conventional intelligent systems, KASERs are capable of accelerated learning in symmetric domains.

Figure 1 plots the knowledge acquired by an intelligent system vs. the cost of acquisition. Conventional expert systems will generate the curve below break-even. That is, with conventional expert systems, cost increases with scale and is never better than linear. Compare this with KASERs where cost decreases with scale and is always better than linear unless the domain has no symmetries (i.e., it is random). Note that such

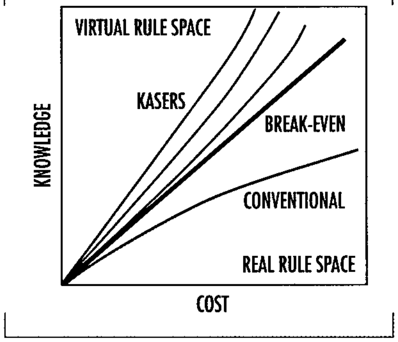


FIGURE 1. The comparative costs of knowledge acquisition.

domains do not exist with scale in practice. Similarly, purely symmetric domains do not exist with scale in practice either. The more symmetric the operational domain, the less the cost of knowledge acquisition and the higher the curve appears in the graph. It is always the case that the virtual rule space >> the real rule space.

INDUCING PROPERTY LISTS

We will define a production system that can automatically acquire a virtual rule space that is exponentially larger than the actual rule space with an exponentially decreasing non-zero likelihood of error. Moreover, the generalization mechanism will not only be bounded in its error, but other than for a straightforward user-query process, it will operate without any *a priori* knowledge supplied by the user.

To begin, define a production rule (e.g., using ANSI Common LISP) to be an ordered pair—the first member of which is a set of antecedent predicates, and the second member of which is an ordered list of consequent predicates. Predicates can be numbers (e.g., [1..2] $\vee$ [10..20]) or words [8].

Previously unknown words or phrases can be recursively defined in terms of known ones. For example, the moves of a Queen in chess (i.e., unknown) can be defined in terms of the move for a Bishop (i.e., known) union those for a Rook (i.e., known). This is a union of property lists. Other basic set operations may likewise be used (e.g., intersection, difference, not, etc.). The use of fuzzy set operators here (e.g., "almost the same as") pertains to computing with words [8].

In a Type I KASER, words and phrases are entered through the use of pull-down menus. In that manner, semantically identical concepts (e.g., Hello and Hi) are not ascribed a distinct syntax, which would otherwise serve to dilute the efficiency of the learning mechanism. In a Type II KASER, distinct syntax may be equated to yield the equivalent normalized semantics. To better visualize this, think of a child who may ask, "What is a bird?" to which the reply is, "It is an animal that flies," to which the question is, "What is an animal?" to which the reply is, "It is a living thing," to which the question is, "What is a living thing?" to which the reply (often) is, "Eat your soup!" (i.e., a Type I delimiter, or stop marker gene).

Two sample rules and their representation follow.

Hydrogen $\wedge$ Oxygen $\wedge$ Spark → Steam

R1: ({Hydrogen, Oxygen, Spark} (Steam))

Hydrogen $\wedge$ Oxygen $\wedge$ Match → Steam

R2: ({Hydrogen, Oxygen, Match} (Steam))

R1 and *R2* may be generalized, since the consequent predicates are identical (i.e., the right-hand sides [RHSs] are equivalent) and the antecedent terms differ in exactly one predicate. This is termed a level-1 generalization because it is one level removed from ground truth. In a level-*i* generalization, *i* is the maximum level of generalization for any antecedent predicate. The need for a generalization squelch arises because contexts may be presented for which there is no matching rule in the real space. Generalizations can be recursively defined.

The advocated approach captures an arbitrary rule's context—something that cannot be accomplished through the use of property lists alone. If veristic terms such as "Warm" are generalized to such terms as "Heat" for example, then qualitative fuzziness will be captured.

A1: ({Heat} {Spark, Match}(X001 Explosive-Gas-Igniter))

Generalization, *A1*, tells us that antecedent predicate, "Heat" is more general than either a Spark or a Match. We may also write this as Heat > {Spark, Match}. Note that the relation ">>" is used to denote ancestral generalizations (and vice versa). The general predicate is initially specified as *X00i*, but this is replaced after interactive query with the user, where possible. Otherwise, the next-level expansions will need to be printed for the user to read. Also, "redundant, at-least-as-specific" rules are always expunged.

The common property list follows the set of instances. Here, the list informs us that a spark or a match may be generalized to Heat because both share the property of being an Explosive-Gas-Igniter. Properties are dynamic. They must be capable of being hierarchically represented, augmented, and randomized. In addition, property lists are subject to set operations (e.g., intersection). Properties can be acquired by way of database and/or user query.

User-queries can be preprocessed by a companion veristic mining system. Similarly, system-generated queries can be post-processed by companion systems. Companion systems can also play a role in imparting tractability to the inference engine.

Consequent terms, being sequences, are taken to be immutable. The idea here is to automatically create a hierarchy of consequent definitions to maximize the potential for rule reuse. Begin by selecting a pair of rules having identical left-hand sides (LHSs), where possible. Consider:

R3: ({Hydrogen, Oxygen, Heat} (Steam))

R4: ({Hydrogen, Oxygen, Heat} (Light, Heat))

Next, an attempt is made to generalize the consequent sequences with the following result.

C1: ((Energy) ((Steam) (Light Heat))(X002 Power-Source))

Here, the properties of Steam intersect those of Light and Heat to yield the property, Power-Source. Thus, a property of Energy, in the current context at least, is that it is a Power Source. Rules *R3* and *R4* are now replaced by their valid generalization, *R5*:

R5: ({Hydrogen, Oxygen, Heat} (Energy))

A key concept is that further learning can serve to correct any latent errors. In addition, notice that as the level of randomization increases on the LHS and RHS, the potential for matching rules, and thus inducing further generalizations, increases by way of feedback. Consequent randomization brings the consequents into a normal form, which then serves to increase the possibility of getting antecedent generalizations, since more RHSs can be equated. Antecedent randomization is similar.

Next, consider *R5*, where *R6* is acquired and appears as follows after substitution using *C1*.

R6: ({Candle, Match} (Energy))

The system always attempts to randomize the knowledge as much as possible. Using *A1* and *C1* leads to the level-1 conjecture, *R7*, which replaces *R6*.

R7: ({Candle, Heat} (Energy))

R7 is not to be generalized with *R6*. This is because {Match, Heat} is the same as {Match, Spark, Match}, which of course reduces to Heat and is already captured by *R7*.

At this point, learning by the system can be demonstrated. Suppose the user asks the system what will happen if a spark is applied to a candle. While this is a plausible method to light a candle, this method will not usually be successful. Thus, the user must report to the system the correct consequent for this action:

R8: ({Candle, Spark} (No-Light))

R8 is a more-specific rule than is *R7* because the former is a level-1 generalization, while the latter is at level-0. Thus, *R8* will be preferentially fired when possible by using a most-specific agenda mechanism. It, too, will be subject to subsequent generalization. Notice that the new consequent will protect against similar error.

The learning process has not completed. We still need to correct the properties list so that Matches and Sparks can be differentiated in the context of lighting a candle. The following property (i.e., LISP) list is obtained.

P1: (Match Explosive-Gas-Igniter Wick-Lighter)

P2: (Spark Explosive-Gas-Igniter)

Now, since Heat is a superclass of Match, its property list is unioned with the new property(s): Wick-Lighter. Suppose, at this point, the user poses the same question, "What will happen if a spark is applied to a candle?" Rule *R7* informs us that it will light; whereas, *R8* informs us that it will not. Again, the inference engine can readily select the appropriate rule to fire because of specialization. However, here there is yet more to learn. Here is what is known: *R7* and *R8* differ on the LHS in exactly one predicate and $prop\ (Energy) \cap prop\ (No - Light) = \varnothing$. The reason that the candle lights for a match, but not for a spark can be delimited by computing, $prop\ (Match) - prop\ (Spark) = prop(P1) - prop(P2) =$ (Wick-Lighter). Rule *R7* is now replaced by *R7'*:

R7': ({Candle, (X003 Wick-Lighter)} (Energy))

that is, a property list named X003 has been substituted for Heat. Notice that X003 is necessarily a subclass of Heat. Then, anything that has (all) the properties on the property list (i.e., X003) can presumably light a candle (e.g., a torch). Observe that the human in the loop need not know why a list of properties is relevant, since the reasons will be automatically discovered. Notice that a Spark can no longer light a candle and only those items having at least Wick-Lighter in their property classes can light a candle. Observe the nonlinear learning that has been enabled here!

Consider now the rule:

R9: ({Candle, Match} (Energy))

Clearly, this rule is correct as written. Candles do indeed produce steam, light, and heat. The usefulness of induction follows from the fact that the system has no knowledge that a candle is a hydrocarbon and hydrocarbons produce steam as a byproduct of combustion.

Antecedent predicate generalizations can be rendered more class-specific as necessary to correct overgeneralizations by increasing the number of levels of available generalization. The rule consequents will not be affected. For example:

A2: ({Car} {Ford, Fiat})

yields:

A3: ({Car} {Family-Car, Sports-Car})

A4: ({Family-Car} {Ford})

A5: ({Sports-Car} {Fiat})

Property lists can be automatically organized into a hierarchical configuration through the use of simple set operations. This means that rules can be generalized or specialized through the use of the disjunctive or conjunctive operators, respectively. Such property lists can be associatively retrieved through the use of a grammatical randomization process [9]. Moreover, matching operations then need to incorporate searching subclasses and superclasses as necessary.

Finally, we note that this system can incorporate fuzzy programming [10]. Fuzzy programming will enable the system to explore a space of alternative contexts as delimited by optional consequent filters and ranked by the level of generalization used to obtain a contextual match (see below).

GRAMMATICAL RANDOMIZATION

Consider the following three property lists:

P3: (Ice A B C)

P4: (Water B C D)

P5: (Steam C E)

Here, ice, water, and steam share a common property, C, which, for example, might be that they are all composed of H_2O. Also, only ice has property A (e.g., frozen); only water has property D (e.g., liquid); and only steam has property E (e.g., gaseous). Observe that only ice and water share property B (e.g., heavier-than-air).

The use of a hierarchical object-representation is fundamental to the specification of property lists, antecedent sets, or consequent sequences. For example, when one specifies the object, "aircraft carrier," one implicitly includes all of its capabilities, subsystems, and the like. One cannot and should not have to specify each subsystem individually. We proceed to develop a randomization for the sample property lists; although, it should be clear that the same approach will work equally well for the antecedent and consequent predicates. Perhaps the most relevant distinction is that one needs to distinguish object sequence dependence from independence in the notation. Of course, property lists are sequence-independent.

As the example stands now, to specify the properties of ice or water, one need state the three properties of each (in any order). This may not seem too difficult, but this is only because the list-size is small. Consider now the randomized version of the property lists:

P3: (Ice A Precipitation)

P4: (Water Precipitation D)

P5: (Steam C E)

P6: (Precipitation B C)

Here, the property of precipitation has been randomized from the property data. Observe, that if the user states property B, then the system will offer the user exactly three choices (e.g., by way of a dynamic pull-down menu): B, Precipitation, or Random. The Random choice allows the user to complete the specification using arbitrary objects. In other words, an associative memory has been defined. Similarly, if the user selects Precipitation, then the system will offer the user exactly four choices (e.g., again by way of a dynamic pull-down menu): Precipitation, Ice, Water, or Random.

Suppose that in keeping with the previously described nomenclature conventions, we had the following property list specifications:

P3': (X004 A Precipitation)

P4': (X005 Precipitation D)

In this case, if the user selects Precipitation, then the system will offer the user the following four choices: Precipitation, A, D, or Random. In other words, it attempts to pattern-match and extrapolate the set.

In practice, randomization is based on known classifications—not arbitrary ones. Thus, in the previous example, the randomization of *P3* and *P4* requires that *P6* be known *a priori.* Again, this still allows for the use of integer identifiers.

Next, it can be seen that the usefulness of randomization is a function of its degree. The relevant question then pertains to how to realize the maximal degree of randomization. First, recall that as rules are generalized, the possibilities for further predicate generalization are increased. This, in turn, implies that the substitution and subsequent refinement of property lists for predicates is increased. Finally, as a result, the virtual space of properly mapped contexts (i.e., conjectures) grows at a rapid rate. Experimental evidence to date indicates that this rate may be exponential for symmetric domains.

Next, we turn our attention to the inference engine, which is common to Type I and II KASERs. Basically, in a Type I KASER, conflict resolution is accomplished through the use of a hierarchical tree of objects evolved by a knowledge engineer, which define generalization and specialization (see below); whereas, in a Type II KASER, conflict resolution is the same as in a Type I KASER, but where the system, instead of the knowledge engineer, evolves hierarchical property lists, which serve to increase the size of the virtual contextual space—without sacrificing convergence in the quality of the response. In effect, declarative knowledge is randomized to yield procedural knowledge.

ACTIVE RANDOMIZATION

Active randomization is a symbiosis of property lists and grammatical randomization. Property lists are really just predicates that are subject to grammatical randomization. Moreover, randomized predicates allow the user to specify contexts and associated actions by using minimal effort [9]. Next, suppose that we had:

(A B C D) (i.e., the properties of A are B, C, D)

Here, A is the randomization of B, C, D. Similarly, we may have

(B E F)

and the two rules (i.e., antecedent differentiation):

R10: A S → W

R11: X S → W

Then, we can create a randomization:

(Q A X)

which, since valid, leads to the following replacement of *R10* and *R11*:

R12: Q S → W

This replacement allows for the possibility of new rule pairings and the desired process then iterates. Thus, we have

(Q: A ∪ X) {expanding A, X}

These are *active transforms* [9] in the sense that whenever A or X change their membership, the properties of Q may change. Evidently, this is a converging process. However, if subsequently we had

R13: A S → T

R14: X S → G

where, T and G have no properties in common (i.e., neither is a subsequence of the other), then it becomes clear that A cannot substitute for X and vice versa. In other words,

R13: (A-X) S → T

R14: (X-A) S → G

Thus, we have

(A: A - X) {contracting A}

(X: X - A) {contracting X}

These are active transforms, and again, this is a converging process. Next, suppose that T and G are such that G is a subsequence of T without loss of generality. Then, it follows that A is a subset of X and

(A: A ∩ X) {contracting A}

(X: X ∪ A) {expanding X}

These are active transforms. This is not, however, necessarily a converging process. That is not to say that it will diverge without bounds. It is just not stable. We do not view this as a problem. It is to be viewed as an oscillatory system that, in some ways, may mimic brain waves. The complexity of interaction will increase as the system is scaled up. The eventual need for high-speed parallel/distributed processing is apparent. The case for consequent differentiation is similar. Here though, one is processing sequences instead of sets.

OBJECT-ORIENTED TRANSLATION MENUS

The Type I KASER requires that declarative knowledge be (dynamically) compiled in the form of object-oriented hierarchical phrase translation menus. Each class (i.e., antecedent and consequent) of bipartite predicates can be interrelated through their relative positions in an object-oriented semantic tree. A declarative knowledge of interrelatedness provides a basis for commonsense reasoning, as will be detailed in the next section. The subject of this section pertains to the creation, maintenance, and use of the object-oriented trees as follows.

1. The phrase-translation menus serve as an intermediate code (as in a compiler) where English sentences can be compiled into menu commands by using rule-based compiler bootstraps. KASERs can be arranged in a network configuration where each KASER can add (post) to or delete from the context of another. This will greatly expand the intelligence of the network with scale and serves to define Minsky's "Society of Mind" [6]. Furthermore, the very-high-level domain-specific language(s) used to define each predicate can be compiled through a network of expert compilers. Alternatively, neural networks can be used to supply symbolic tokens at the front end.
2. Each antecedent or consequent phrase can be associated with a textual explanation, Microsoft's Text-to-Speech engine (Version 4.0), an audio file, a photo, and/or a video file. Images may be photographs, screen captures, scans, drawings, etc. They may also be annotated with arrows, numbers, etc. Voice navigation may be added at a later date.
3. Antecedents and consequents can be captured by using an object-oriented approach. The idea is to place descriptive phrases in an object-oriented hierarchy such that subclasses inherit all of their properties from a unique superclass and may include additional properties as well. Menus can beget submenus, and new phrases can be acquired at any level.

   Consider the partial path, office supply, paper clip and the partial path, conductor, paper clip. Here, any subclass of paper clip will have very different constraints depending on its derivation. For example, anything true of paper clips in the context of their use as conductors must hold for every subclass of paper clips on that path. Unique antecedent integers can be set up to be triggered by external rules. Similarly, unique consequent integers can be set up to fire external procedures. All we need do is facilitate such hooks for future expansion (e.g., the radar-mining application domain).

   Each project is saved as a distinct file, which consists of the antecedent and consequent trees, the associated rule base, and possibly the multimedia attachments.
4. A tree structure and not a graph structure is appropriate because the structure needs to be readily capable of dynamic acquisition (i.e., relatively random phrases) and deletion, which cannot be accomplished in the presence of cycles due to side effects. Note that entering a new phrase in a menu implies that it is semantically distinct from the existing phrases, if any, in that menu.
5. A tree structure is mapped to a context-free grammar (CFG), where the mapping process needs to be incremental in view of the large size of the trees. Each node or phrase is assigned a unique number, which serves to uniquely identify the path.
6. Each phrase may be tagged with a help file, which also serves the purposes of the explanation subsystem. This implies that conjuncts are not necessary to the purpose of the antecedent or consequent trees.

7. Each menu should be limited to on the order of one screen of items (e.g., 22). Toward this end, objects should be dynamically subdivided into distinct classes. That is, new submenus can be dynamically created and objects moved to or from them.
8. Three contiguous levels of hierarchy should be displayed on the graphical user interface (GUI) at any time, if available.
9. A marker gene or bookmark concept allows the user to set mark points for navigational purposes.
10. A list of recently visited menus serves to cache navigational paths for reuse.
11. A global find mechanism allows the user to enter a phrase and search the tree from the root or present location and find all matches for the phrase up to a prespecified depth. The path, which includes the phrase, if matched, is returned.
12. Entered phrases (i.e., including pathnames) can be automatically extrapolated where possible. This "intellisense" feature facilitates keyboard entry. It can also assist with the extrapolation of pathnames to facilitate finding or entering a phrase. Pathname components may be truncated to facilitate presentation.
13. A major problem in populating a tree structure is the amount of typing involved. In view of this, copy, paste, edit, and delete functions are available to copy phrases from one or more menus to another through the use of place-holding markers. Phrase submenus are not copied over because distinct paths tend to invalidate submenu contents in proportion to their depth. Again, new integers are generated for all phrases. Note that the returned list of objects still needs to be manually edited for error and/or omissions. This follows from randomization theory. This maps well to natural language translation.
14. Disjuncts in a menu serve as analogs and superclasses serve as generalizations for an explanation subsystem. In addition, help files and pathnames will also serve for explanative purposes.
15. An "intellassist" feature allows the system to predict the next node in a contextual, antecedent, or consequent tree. Each node in a tree locally stores the address (number) of the node to be visited next in sequence. If a node has not been trained, or if the pointed-to address has been deleted without update, then a text box stating "No Suggestion" pops up, and no navigation is effected if requested. Otherwise, potentially three contiguous menus are brought up on the screen, where the farthest right menu contains the addressed node. Navigation is accomplished by clicking on a "Suggest" button. Otherwise, all navigation is manually performed by default. The user can hop from node to node by using just the suggest button without registering an entry. The use of any form of manual navigation enables a "Remember" button immediately after the next term, if any is entered. Clicking on this enabled button will result in setting the address pointed to by the *previously entered* node to that of the *newly entered* node. The old pointer is thus overwritten. Note that this allows for changing the item selected within the same menu. Note, too, that if a node (e.g., Toyota) is deleted, then all pointers to it may be updated to the parent class (e.g., car menu) and so on up the tree (e.g., vehicle type menu).

A pull-down menu will enable one of two options: (1) Always Remember (by default) and (2) Remember when Told. The Remember button is not displayed under option (1), but the effect under this option is to click it whenever it would have otherwise been enabled. The system always starts at the root node of the relevant tree.

16. It does not make sense to retain a historical prefix for use by the intellassist feature. That is, there is no need to look at where you were to determine where you want to go. While potentially more accurate, this increase in accuracy is more than offset by the extra training time required, the extra space required, and the fact that it will take a relatively long time to reliably retrain the nodes in response to a dynamic domain environment.

AN A* ORDERED SEARCH ALGORITHM

Expert compilers apply knowledge bases to the effective translation of user-specified semantics [11]. The problem with expert compilers is that they use conventional expert systems to realize their knowledge bases. A KASER is advocated because it can amplify a knowledge base by using an inductively extensible representational formalism.

Here, we present a relatively high-level view of the KASER algorithm. We claim that it represents a great advance in the design of intelligent systems by reason of its capability for symbolic learning and qualitative fuzziness:

1. Click on antecedent menus to specify a contextual conjunct. Alternatively, a manual "hot button" will bring up the immediately preceding context for reuse or update. Renormalization is only necessary if a generalization was made—not for term deletion (see below). Iteratively normalize the context (i.e., reduce it to the fewest terms) by using the tree grammar. Note that contextual normalization can be realized in linear time in the number of conjuncts and the depth of search. Here are the reduction rules, which are iteratively applied in any order—allowing for concurrent processing:
   a. S → A | B | C ... then replace A, B, C ... with S just in case all of the RHS is present in the context. This step should be iteratively applied before moving on to the next one.
   b. S → A ... and A → B ... and B → C ... then if S, A, B, C are all present in the context, then remove A, B, C since they are subsumed by S. It is never necessary to repeat the first step after conclusion of the second.
2. Compute the specific stochastic measure. Note that the specific stochastic measure does not refer to validity—only to the creative novelty relative to the existing rules while retaining validity. For example, given the antecedent grammar: C5 → C3 | C4; C4 → C1 | C2:
   a. {C3 C1} {{C3}, {C2 C3},} covers and matches the first {C3} at level 0. Note that the first covered match, if any, that does not have a covered superset is the one to be fired—a result that follows from the method of transposition.
   b. {C3 C1} {{C5}, {C2 C3},} matches nothing at the level 0 expansion, so we expand the RHS with the result, {C3 C1} {{*C5 (C3 C4)},

{*C2 *C3},} where the C2 C3 are both primitives and *Ci can be matched, but not expanded again. (..) is used to denote disjunction. Here, {C5} is matched at level 1. Note that at any level, only one term inside the parentheses (e.g., C3) need be covered to get a match of any one disjunct.

c. {C3 C6} {{C2}, {C5 C6},} matches nothing at the level 0 expansion, so we expand the RHS with the result, {C3 C6} {{*C2}, {*C5 (C3 C4), *C6}}, which matches at level 1 because we matched (C3 OR C4) AND C6. Note that C6 was never expanded because it was pre-matched by the existing context. This economy is possible as a result of pre-normalizing the context.

d. The result of applying the method of transposition to the above step is {{C5 C6}, {C2},}.

e. Each matched {...} fires a consequent, which, if not primitive, matches exactly one row header (i.e., a unique integer) and step (2) iterates.

f. Maintain a global sum of the number of levels of expansion for each row for each consequent term. The specific stochastic measure is taken as the maximum of the number of levels of expansion used for each consequent term.

3. Exit the matching process with success (i.e., for a row) or failure based on reaching the primitive levels, a timer-interrupt, a forced interrupt, and/or by using the maximum allocated memory.

4. If a sequence of consequent actions has been attached, then the sequence is pushed onto a stack in reverse order such that each item on the stack is expanded in a depth-first manner. A parenthesized sequence of actions will make clear the hierarchy. For example, ((Hold Writing Instrument (Hold Pencil with Eraser)) (Press Instrument to Medium (Write Neatly on Paper))). Here, the subclasses are nested. Such a representation also serves explanative purposes. Thus, here one has, Hold Writing Instrument, Press Instrument to Medium, at the general level, and Hold Pencil with Eraser, Write Neatly on Paper, at the specific level. A companion intelligent system could transform the conceptual sequences into smooth natural language (e.g., Pick up a pencil with an eraser and write neatly on a sheet of paper.) Set the general stochastic measure (GSM) to zero. Note that the stochastic measures for each predicate are computed and held in a data structure. The data will be used by the inference engine.

5. If a match is not found, then since we already have an expanded antecedent {...}, we proceed to expand the context in a breadth-first manner (i.e., if enabled by the level of permitted generalization). Compute the general stochastic measure. Initialize the general stochastic measure to GSM. Note that the general stochastic measure is a measure of validity. Set the starting context to the context.

a. A specialized match was sought in step (2), and a generalized match is sought here. Expanding the context can lead to redundancies. For example, {*C1 *C2 *C3 *C4 C1 C2}. Here, the solution is to simply not include any term that is already in the (expanded) context. Stochastic accuracy is thus preserved. Any method that does not preserve stochastic accuracy is not to be used.

b. {C5} {{*C3}, {*C2 *C3},} failed to be matched in step (2), so a level 1 expansion of the context is taken:

 {*C5 C3 C4} {{*C3}, {*C2 *C3},} where C3 is matched at level 1.

c. {C5 C6} {{*C1}, {*C2 *C3},} matches nothing at level 0, so a level 1 expansion of the context is taken:

 {*C5 C3 C4 *C6} {{*C1}, {*C2 *C3},} matches nothing at level 1, so a level 2 expansion of the context is taken:

 {*C5 *C3 *C4 C1 C2 *C6} {{*C1}, {*C2 *C3},} matches C1 OR C2 AND C3 at level 2. The first covered set is the one to be fired (i.e., even though both sets are covered), since it does not have a covered superset. Next, the method of transposition is trivially executed with no resulting change in the logical ordering.

d. Each matched {...} fires a consequent, which, if not primitive, matches exactly one row header (i.e., a unique integer) and step (2) iterates.

e. One should maintain a count of the maximum number of levels of expansion for the context below the initial level. The general stochastic measure is defined by GSM plus the maximum number of levels that the context minimally needs to be expanded to get the "first" (i.e., method of transposition) match. This stochastic is represented by the maximum depth for any expansion.

f. If the context fails to be matched, then generalize each term in the starting context one level up in the tree. Remove any redundancies from the resulting generalization. If the generalized context differs from the starting context, then add one to GSM and go to step (5). Otherwise, go to step (6). For example, the starting context {C2 C3} is generalized to yield {C4 C5}. If this now covers a {..}, then the general stochastic measure is one. Otherwise, it is subsequently expanded to yield {*C4 C1 C2 *C5 C3 C4} at the first level. Notice that the second C4 has a longer derivation, is redundant, and would never have been added here. Note, too, that C4 is also a sibling or analog node. If this now covers a {..}, then the GSM remains one, but the specific stochastic measure is incremented by one to reflect the additional level of specialization.

 For another example, Toyota and Ford are instances of the class car. If Toyota is generalized to obtain car, which is subsequently instantiated to obtain Ford (i.e., an analog), then the general and specific stochastic measures would both be one. The general stochastic measure represents the number of levels of expansion for a term in one direction, and the specific stochastic measure represents the number of levels of expansion from this extrema in the opposite direction needed to get a match. The final general (specific) stochastic is taken as the maximum general (specific) stochastic over all terms.

g. Conflict resolution cannot be a deterministic process as is the case with conventional expert systems. This is because the number of predicates in any match must be balanced against the degree of specialization and/or generalization needed to obtain a match. Thus, a heuristic approach is required. The agenda mechanism will order the rules by their size, general stochastic, and specific stochastic with recommended weights of 3, 2, and 1 respectively.

6. Exit the matching process with success (i.e., for the entire current context for a row) or failure based on reaching the primitive levels, a timer-interrupt, a forced interrupt, and/or by using the maximum allocated memory. Note that a memory or primitive interrupt will invoke step (5f). This enables a creative search until a solution is found or a timer-interrupt occurs. Note, too, that it is perfectly permissible to have a concept appear more than once for reasons of economy of reference, or to minimize the stochastic measures (i.e., provide confirming feedback). The stochastic measures also reflect the rapidity with which a concept can be retrieved.
7. Knowledge acquisition:
   a. Note that new rules are added at the head.
   b. If exit occurs with failure, or the user deems a selected consequent (e.g., in a sequence of consequents) in error (i.e., trace mode on), then the user navigates the consequent menus to select an attached consequent sequence, which is appropriate for the currently normalized context.
   c. If the user deems that the selected "primitive" consequent at this point needs to be rendered more specific, then a new row is opened in the grammar, and the user navigates the consequent menus to select an attached consequent sequence.
   d. A consequent sequence can pose a question, which serves to direct the user to enter a more specific context for the next iteration (i.e., conversational learning). Questions should usually only add to the context to prevent the possibility of add/delete cycles.
   e. Ask the user to eliminate as many specific terms (more general terms will tend to match more future contexts) from the context as possible (i.e., and still properly fire the selected consequent sequence given the assumptions implied by the current row). A context usually consists of a conjunct of terms. This tends to delimit the generality of each term as it contributes to the firing of the consequent. However, once those antecedent terms become fewer in number for use in a subsequent row, then it becomes possible to generalize them while retaining validity. The advantage of generalization is that it greatly increases reusability. Thus, we need to afford the user the capability to substitute a superclass for one or more terms. Note that this implies that perfectly valid rules that were entered can be replayed with specific (not general) stochastics greater than zero. This is proper, since the specific stochastic preserves validity in theory. Thus, the user may opt to generalize one or more contextual terms by backtracking their derivational paths. If and only if this is the case, step (1) is applied to normalize the result. An undo/redo capability is provided. Validated rule firings are only saved in the rule base if the associated generalization stochastic is greater than zero. The underlying assumption is that rule instances are valid. If a pure rule instance proves to be incorrect, then the incorrect rule needs to be updated or purged, and the relevant object class menu(s) may be in need of repair. For example, what is the minimal context to take FIX_CAR to FIX_TIRE? A companion intelligent system could learn to eliminate and otherwise generalize specific terms (e.g., randomization theory).

f. The system should verify for the user all the other {...} in the current row that would fire or be fired by the possibly overgeneralized {...} if matched. (Note that this could lead to a sequence of UNDOs.)

   For example, ({C5} A2) {({C5} A1) ({C5 C6} A2) ({C5 C7} A2, A3)} informs the user that if the new C5 acquisition is made, then A2 and not A1 is proper to fire. If correct, then the result is {{C5} A2 {C5 C7} A2, A3}. {C5 C6} A2 has been eliminated because it is redundant. Also, {C5 C7} A2, A3 is fired just in case C5 AND C7 are true—in which case, it represents the most specific selection since it is a superset of the first set. If the elimination of one or more specific terms causes one or more {...} to become proper supersets, then warning message(s) may be issued to enable the user to reflect on the proposed change(s). If the elimination and/or generalization of one or more specific terms enables the firing of another rule in the same row in preference to the generalized rule, then the generalization is rejected as being too general. Note that there is no need to normalize the results, as they would remain in normal form. Also, any further normalization would neutralize any necessary speedup.

g. A selected consequent number may not have appeared on the trace path with respect to the expansion of each consequent element taken individually. Checking here prevents cycle formation.

h. It should never be necessary to delete the least frequently used (LFU) consequent {...} in view of reuse, domain specificity, processor speed, and available memory relative to processor speed. Nevertheless, should memory space become a premium, then a hierarchy of caches should be used to avoid deletions.

8. A metaphorical explanation subsystem can use the antecedent/consequent trees to provide analogs and generalizations for explanative purposes. The antecedent/consequent paths (e.g., ROOT, FIX_CAR, FIX_TIRE, etc.) serve to explain the recommended action in a way similar to the use of the antecedent and consequent menus. The antecedent/consequent menus will provide disjunction and "user-help" to explain any level of action on the path. Note that the system inherently performs a fuzzy logic known as computing with words [4] (i.e., based on the use of conjuncts, descriptive phrases, and tree structures). The virtual rule base is exponentially larger than the real one and only limited by the number of levels in the trees, as well as by space-time limitations on breadth-first search imposed by the hardware.

9. A consequent element could be a "do-nothing" element if need be (i.e., a Stop Expansion). The provision for a sequence of consequents balances the provision for multiple antecedents. The selected consequent(s) need to be as general class objects as can be to maximize the number of levels and, thus, the potential for reuse at each level. The consequent grammar is polymorphic since many such grammars can act (in parallel via the Internet) on a single context with distinct, although complementary results. Results can be fused as in a multilevel, multicategory associative memory. Multiple context-matched rules may not be expanded in parallel because there can be no way to ascribe *probabilities* to partially order the competing rules and

because any advantage would be lost to an exponential number of context-induced firings. The consequent {...}s cannot be ranked by the number of matching terms (i.e., for firing the most specific first) because the most specific terms are generally incomparable. However, a covered superset is always more specific than any of its proper subsets. Thus, the first covered set that does not have a covered superset in the same row is the one to be fired. If it does have a covered superset, then the superset is fired only if it is the next covered one to be tested in order. It is not appropriate to tag nodes with their level, use a monotonically increasing numbering system, or any equivalent mechanism to prevent the unnecessary breadth-first expansion of a node(s) because the menus are dynamic, and it would be prohibitively costly to renumber, for example, a terabyte of memory. Note that node traversal here is not synonymous with node visitation. Even if parallel processors could render the update operation tractable, the search limit would necessarily be set to the depth of the deepest unmatched node. Here, the likelihood of speedup decreases with scale. The contextual terms should only be *'d if this does not interfere with their expansion—even if normalized. Let the context be given as {C5 C6} and the RHS be {C5 C7}, {C1 ...},. Clearly, if the context had *C5, then the C1 might never be matched.

10. Unlike the case for conventional expert systems, a KASER cannot be used to backtrack consequents (i.e., goal states) to find multiple candidate antecedents (i.e., start states). The problem is that the preimage of a typical goal state cannot be effectively constrained (i.e., other than for the case where the general and specific stochastics are both zero) in as much as the system is qualitatively fuzzy. Our answer is to use fuzzy programming in the forward-chained solution. This best allows the user to enter the constraint knowledge that he/she has into the search. For example, if the antecedent menus are used to specify CAR and FUEL for the context and the consequent is left unconstrained for the moment, then the system will search through all instances, if any, of CAR crossed with all instances of FUEL (i.e., to some limiting depth) to yield a list of fully expanded consequents. Generalization-induced system queries, or consequents that pose questions, if any, will need to be answered to enable this process to proceed. Thus, in view of the large number of contexts that are likely to be generated, all interactive learning mechanisms should be disabled or bypassed whenever fuzzy programming is used. Note that CAR and FUEL are themselves included in the search. Each predicate can also be instantiated as the empty predicate in the case of the antecedent menus, if user-enabled. If the only match occurs for the case of zero conjuncts, then the consequent tree is necessarily empty. A method for fuzzy programming is to simply allow the user to split each conjunct into a set of disjuncts and expand all combinations of these to some fixed depth to obtain a list of contexts. This use of a keyword filter, described below, is optional. For example, the specification $(A \vee A' \vee !A'') \wedge (B \vee B') \wedge (C)$ yields 23 candidate contexts—including the empty predicate (i.e., if one assumes that A'' is primitive and allows for redundancy), which excludes the empty context. The exclamation mark, "!", directs the system to expand the nonterminal that follows it to include (i.e., in addition to itself) all of the next-level instances of its class. For example, !CAR would yield (CAR TOYOTA FORD MAZDA HONDA ... λ). Here, lambda denotes the empty predicate and is included as a user option.

A capability for expanding to two or more levels if possible (e.g., "!!") is deemed to be nonessential but permissible (e.g., for use with relatively few conjuncts). This follows because the combinatorics grow exponentially. One can always take the most successful context(s) produced by a previous trial, expand predicates to another level by using "!s" where desired, and rerun the system. Note that, in this manner, the user can insert knowledge at each stage—allowing for a far more informed, and thus, deeper search than would otherwise be possible. Moreover, the fuzzy specialization engine will stochastically rank the generalized searches to enable an accurate selection among contexts for possible rerun.

The search may be manually terminated by a user interrupt at any time. The search is not to be automatically terminated subsequent to the production of some limit of contexts because to do so would leave a necessarily skewed distribution of contexts—thereby giving the user a false sense of completeness. We would rather have the user enter a manual interrupt and modify the query subsequently. A terminated search means that the user either needs to use a faster computer, or more likely, just narrow down the search space further and resubmit. For example, if we have the antecedent class definitions:

(CAR (FORD TOYOTA)) (FUEL (REGULAR_GAS HIGH_TEST DIESEL)) (AGE (OLD (TIRES ...)) (NEW (TIRES ...)))

and the contextual specification:

(!CAR) ∧ (!FUEL) ∧ (NEW),

then we would have the following 35 contexts allowing for the empty predicate. Note that the use of the empty predicate is excluded by default, since its use is associated with an increase in the size of the search space and since it may not be used with the consequent menus (see below).

CAR
DIESEL
FORD
FUEL
HIGH_TEST
NEW
REGULAR_GAS
TOYOTA
CAR DIESEL
CAR FUEL
CAR HIGH_TEST
CAR NEW
CAR REGULAR_GAS
DIESEL NEW
FORD NEW
FUEL NEW
HIGH_TEST NEW
REGULAR_GAS NEW
TOYOTA FUEL
TOYOTA DIESEL
TOYOTA HIGH_TEST
TOYOTA NEW
TOYOTA REGULAR_GAS
CAR DIESEL NEW

CAR FUEL NEW
CAR HIGH_TEST NEW
CAR REGULAR_GAS NEW
FORD DIESEL NEW
FORD FUEL NEW
FORD HIGH_TEST NEW
FORD REGULAR_GAS NEW
TOYOTA DIESEL NEW
TOYOTA FUEL NEW
TOYOTA HIGH_TEST NEW
TOYOTA REGULAR_GAS NEW

The user may also have used the consequent menus to specify an optional conjunctive list of key phrases, which must be contained in any generated consequent. Those generated consequents, which contain the appropriate keywords or phrases, are presented to the user in rank order—sorted first in order of increasing generalization stochastic and within each level of generalization stochastic in order of increasing specialization stochastic (i.e., best-first). For example, (general, specific) (0, 0) (0, 1) (1, 0) (1, 1) ... Recall that only the specific stochastic preserves validity.

The specified antecedent and consequent classes should be as specific as possible to minimize the search space. Neither the antecedent nor consequent terms specified by the user are ever generalized. For example, if we have the consequent class definitions:

(COST_PER_MILE (CHEAP MODERATE EXPENSIVE))
(MPG (LOW MEDIUM HIGH))

then we can constrain the space of generated consequents in a manner similar to the way in which we constrained the space of generated antecedents. Thus, for example we can write:

(!CAR) ∧ (!FUEL) ∧ (NEW) —> (!COST_PER_MILE) ∧ (!MPG)

This is orthogonal programming; that is, reusing previous paradigms unless there is good reason not to reuse them. Each candidate solution has been constrained so that it must contain at least one phrase from the four in the COST_PER_MILE class *and* at least one phrase from the four in the MPG class—including the class name, but excluding the empty predicate of course. IF an asterisk, "*" is placed after the arrow, then the compiler is directed not to filter the produced consequents in any way.

The user can make changes wherever (i.e., to the antecedents, the consequents, or both) and whenever (e.g., interactively) appropriate and rerun the system query. This represents *computing with words* because fuzziness occurs at the qualitative level. It is not really possible for distinct classes to produce syntactically identical phrases because pathnames are captured using unique identifiers. That is, the identifiers are always unique even if the represented syntax is not.

It is not necessary to weight the consequent phrases because instance classes preserve validity (i.e., at least in theory) and because it would be otherwise impossible to ascribe weights to combinations of words or phrases. For example, "greased" and "lightning" might be synonymous with fast, but taken together (i.e., "greased lightning"), an appropriate weight should be considerably greater than the sum of the partial weights. The degree to which the conjunctive weight should be increased does not lend itself to practical determination. Moreover, one is then faced with the indeterminable question (i.e., for ranking) as to which is the more significant metric: the weight or the two stochastics. Besides, if one follows the dictates of quantum mechanics or veristic computing, it suffices to rank consequent phrases by group as opposed to individually.

Feedback produced, in the form of implausible generalizations, serves to direct the knowledge engineer to modify the involved declarative class structures by regrouping them into new subclasses so as to prevent the formation of the erroneous generalizations. This, too, is how the system learns. The iterative pseudocode for accomplishing the combinatorial expansion follows.

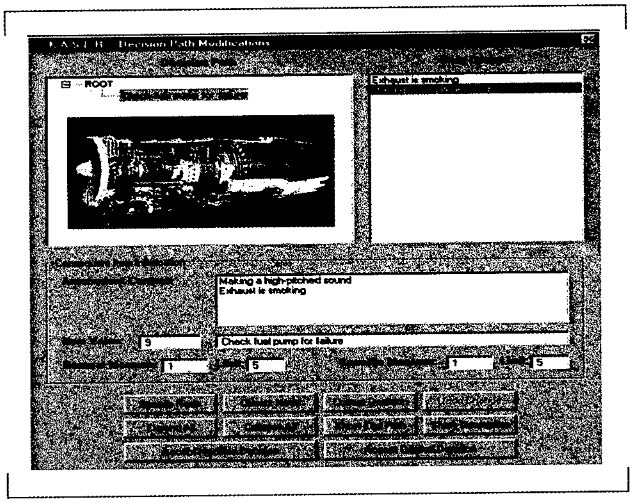
```
1. Initialize the list of Candidate Contexts to $\lambda$.
2. Each conjunct—e.g., $(A \vee A' \vee !A'')$—in the starting list—e.g., $(A \vee A' \vee !A'') \wedge (B \vee B') \wedge (C)$ will be processed sequentially.
3. Note that $!A''$ means to expand the disjunct to include all members of its immediate subclass, if any. Similarly, $!!A''$ means to expand the disjunct to a depth of two. The provision for multilevel expansion is implementation-dependent and is thus optional. Each expanded conjunct is to be augmented with exactly one $\lambda$ if and only if the user has enabled the $\lambda$-option. This option is disabled by default.
4. Expand the first conjunct while polling for a manual interrupt. Here, the result is
   $(A \vee A' \vee A'' \vee A''.a \vee A''.b \vee \lambda)$ .
5. Note that the fully expanded list of conjuncts for illustrative purposes appears:
   $(A \vee A' \vee A'' \vee A''.a \vee A''.b \vee \lambda) \wedge$
   $(B \vee B' \vee \lambda) \wedge (C \vee \lambda)$
6. Initialize a buffer with the disjuncts in the first conjunct. Here, the first six buffer rows are populated.
7. Copy the contents of the buffer to the top of the list of Candidate Contexts;
8. Current Conjunct = 2;
9. Note that there are three conjuncts in this example.
10. WHILE (Current Conjunct <= Number of Conjuncts) and NOT Interrupt DO
    {
11. Expand the Current Conjunct while polling for a manual interrupt.
12. Let d = the number of disjuncts in the Current Conjunct;
13. Using a second buffer, duplicate the disjuncts already in the first buffer d times. For example, here, the second conjunct has three disjuncts and would thus result in the buffer: A, A, A, A', A', A', A'', ... , $\lambda$ , $\lambda$, $\lambda$.
14. FOR each element i in the buffer WHILE NOT Interrupt DO
15.     FOR each Disjunct j in the Current Conjunct WHILE NOT Interrupt DO
        {
16.         Buffer [i] = Buffer [i] || Current Disjunct [j].
17.         (For example, AB, AB', A $\lambda$, A'B, A'B', A' $\lambda$, ..., $\lambda$B, $\lambda$B', $\lambda\lambda$.)
        }
18. IF the $\lambda$-option has been enabled THEN
        Append the contents of the buffer to the bottom of the list of Candidate Contexts while polling for a manual interrupt.
19. Current Conjunct++
    }
20. An interrupt may be safely ignored for the next two steps.
21. IF the $\lambda$-option has been enabled THEN
       Final Contexts = Candidate Contexts - $\lambda$
22. ELSE
       Final Contexts = contents of the buffer.
23. Duplicate contexts are possible due to the use of $\lambda$ and possible duplicate
    entries by the user. Searching to remove duplicate rows is an
    $O(n^2)$ process. Thus, it should never be mandated, but rather offered
    as an interruptible user-enabled option.
```

The iterative pseudocode for constraining the generated consequents follows.

```
1. Expand each conjunct—e.g., $(A \vee A' \vee !A'')$—in the starting list—e.g.,
   $(A \vee A' \vee !A'') \wedge (B \vee B') \wedge (C)$. Note that the $\lambda$-option is disabled.
2. Here, the result is
   $(A \vee A' \vee A'' \vee A''.a \vee A''.b) \wedge (B \vee B') \wedge (C)$.
3. FOR each consequent sequence (i.e., rule) WHILE NOT Interrupt
   DO
   {
4.     match = FALSE;
5.     FOR each expanded conjunct (i.e., required key concept)
       WHILE NOT Interrupt DO
       {
6.         FOR each predicate in an expanded conjunct (i.e., PEC)
           WHILE NOT Interrupt DO
           {
7.             FOR each predicate in a consequent sequence (i.e.,
               PICS) WHILE NOT Interrupt DO
               {
8.                 IF PEC = PICS THEN
                   {
                       match = TRUE;
                       BREAK;
                       BREAK;
                       (Each BREAK transfers
                       control to the next statement
                       outside of the current loop.)
                   }
               }
           }
9.         IF NOT match THEN BREAK
       }
10.    IF NOT match THEN remove current rule
       from the candidate list
11.    ELSE the rule is saved to the set of candidate
       rules, which is sorted as previously described.
   }
```

SUGGESTED NAVAL APPLICATIONS

Figure 2 presents a screen capture of a Type I KASER for diagnosing faults in a jet engine. Observe that the general and specific stochastics are both one. This means, in the case of the general stochastic, that the KASER needed to use a maximum of one level of inductive inference to arrive at the prescribed action. Similarly, the specific stochastic indicates that a maximum of one level of deduction was necessarily employed to arrive at this prescribed action. Contemporary expert systems would not have been able to make a diagnosis and prescribe a course of action, since they need to be explicitly programmed with the necessary details. In other words, the KASER is offering a suggestion here that is open under deductive process. Simply put, it created new and presumably correct knowledge. Here are the two level-0 rules, supplied by the knowledge engineer (i.e., *R15* and *R16*), that were used in conjunction with the declarative object trees to arrive at the new knowledge, *R18*:



Exhaust is smoking

Making a high-pitched sound
Exhaust is smoking

9

Check fuel pump for failure

1 5 1 5

FIGURE 2. Screen capture of an operational Type I KASER.

R15: If Exhaust Flaming and Sound Low-Pitched Then Check Fuel Injector for Carbonization

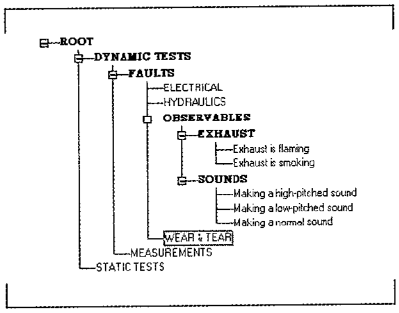
R16: If Exhaust Smokey and Sound High-Pitched Then Check Fuel Pump for Failure

R17: If Exhaust Smokey and Sound Low-Pitched Then Check Fuel Pump for Failure

Upon confirmation of *R17*, *R16* and *R17* are unified as follows.

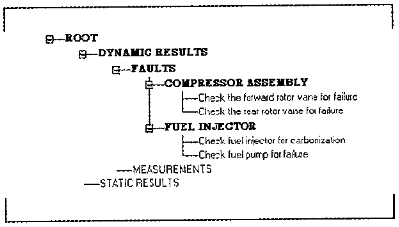
R18: If Exhaust Smokey and Sound Not Normal Then Check Fuel Pump for Failure

The KASER finds declarative antecedent knowledge, which informs the system that the three sounds that an engine might make, subject to dynamic modification, are high-pitched, low-pitched, and normal. By generalizing high-pitched sounds one level to SOUNDS (see Figure 3) and then specializing it one level, one arrives at the first-level analogy: low-pitched sounds. This analogy enables the user context to be matched and leads to the creation of new knowledge. Figure 4 depicts the consequent tree and is similar to the antecedent tree shown in Figure 3. The consequent tree is used to generalize rule consequents so as to maximize reusability. Object reuse may simultaneously occur at many levels, even though this example depicts only one level for the sake of clarity. There are many more algorithms, settings, and screens that may be detailed.



```
ROOT
    DYNAMIC TESTS
        FAULTS
            ELECTRICAL
            HYDRAULICS
            OBSERVABLES
                EXHAUST
                    Exhaust is flaming
                    Exhaust is smoking
                SOUNDS
                    Making a high-pitched sound
                    Making a low-pitched sound
                    Making a normal sound
            WEAR & TEAR
        MEASUREMENTS
    STATIC TESTS
```

FIGURE 3. Screen capture of an antecedent tree.



```
ROOT
    DYNAMIC RESULTS
        FAULTS
            COMPRESSOR ASSEMBLY
                Check the forward rotor vane for failure
                Check the rear rotor vane for failure
            FUEL INJECTOR
                Check fuel injector for carbonization
                Check fuel pump for failure
        MEASUREMENTS
    STATIC RESULTS
```

FIGURE 4. Screen capture of a consequent tree.

Another application is the automatic classification of radar signatures. Basically, the radar data are assigned a feature set in consultation with an expert. Next, a commercial data-mining tool is applied to the resulting very large database to yield a set of rules and associated statistics. These rules are manually fed into the Type I KASER, which interacts with the knowledge engineer to create the antecedent and consequent trees, as well as a fully generalized rule base and miscellaneous sundry. Upon completion of the manual acquisition, the KASER is given a procedure to link it through open database connectivity (ODBC) to an external electronic intelligence (ELINT) database. This database supplies the radar signatures in approximately real time. The signatures are then automatically classified by the KASER's virtual rule space and the generated stochastics provide an indication of reliability. The KASER, having a virtual rule space >> real rule space can produce erroneous advice if the general stochastic is greater than zero. In this event, the user is requested to supply a corrective consequent(s), which may be "radioed" to the base computer for subsequent update on a daily basis, followed by uploading the more learned KASER. The main benefit here is that the KASER can supply solutions to complex signature-identification problems that would not be cost-effective to supply otherwise (see Figure 1). A Type II KASER should be able to automatically acquire the feature set.

CONCLUSIONS

This project seeks to demonstrate (1) a strong capability for symbolic learning, (2) an accelerating capability to learn, (3) conversational learning (i.e., learning by asking appropriate questions), (4) a metaphorical explanation subsystem, (5) probabilistically ranked alternative courses of action that can be fused to arrive at a consensus that is less sensitive to occasional errors in training, and (6) a capability to enunciate responses. It is argued that the intelligent components of any Command Center of the Future (CCOF) cannot be realized in the absence of a strong capability for symbolic learning.

Randomization theory holds that the human should supply novel knowledge exactly once (i.e., random input), and the machine should extend that knowledge by way of capitalizing on domain symmetries (i.e., expert compilation). In the limit, novel knowledge can only be furnished by chance itself. This means that, in the future, programming will become more creative and less detailed, and thus, the cost per line of code will rapidly decrease. According to Bob Manning [12]: "Processing knowledge is abstract and dynamic. As future knowledge management applications attempt to mimic the human decision-making process, a language is needed that can provide developers with the tools to achieve these goals. LISP enables programmers to provide a level of intelligence to knowledge-management applications, thus enabling ongoing learning and adaptation similar to the actual thought patterns of the human mind."

Moreover, according to Erann Gat at the Jet Propulsion Laboratory, California Institute of Technology, working under a contract with the National Aeronautics and Space Administration [13]: "Prechelt concluded that 'as of JDK 1.2, Java programs are typically much slower than programs written in C or C++. They also consume much more memory.' "

Gat states that "We repeated Prechelt's study by using Franz Inc.'s Allegro Common LISP 4.3 as the implementation language. Our results show that LISP's performance is comparable to or better than C++ in execution speed; it also has significantly lower variability, which translates into reduced project risk. The runtime performance of the LISP programs in the aggregate was substantially better than C and C++ (and vastly better than Java). The mean runtime was 41 seconds versus 165 for C and C++. Furthermore, development time is significantly lower and less variable than either C++ or Java. This last item is particularly significant because it translates directly into reduced risk for software development.

Memory consumption is comparable to Java. LISP thus presents a viable alternative to Java for dynamic applications where performance is important."

In conclusion, the solution to the software bottleneck will be cracking the knowledge-acquisition bottleneck in expert systems (compilers).

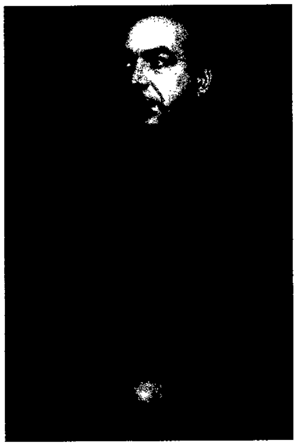
ACKNOWLEDGMENTS

I would like to thank Robert Rush, Jr., and James Boerke for their technical programming support in the implementation of the Type I KASER. The Office of Naval Research sponsored this In-house Laboratory Independent Research project.

REFERENCES

1. Chaitin, G. J. 1975. "Randomness and Mathematical Proof," *Scientific American*, vol. 232, no. 5, pp. 47–52.
2. Uspenskii, V. A. 1987. *Gödel's Incompleteness Theorem*, translated from Russian. Ves Mir Publishers, Moscow, Russia.
3. Lin, J-H. and J. S. Vitter. 1991. "Complexity Results on Learning by Neural Nets," *Machine Learning*, vol. 6, no. 3, pp. 211–230.
4. Rubin, S. H. 1999. "Computing with Words," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 29, no. 4, pp. 518–524.
5. Feigenbaum, E. A. and P. McCorduck. 1983. *The Fifth Generation*, Addison–Wesley Publishing Company, Reading, MA.
6. Minsky, M. 1987. *The Society of Mind*. Simon and Schuster, Inc., New York, NY.
7. Clark, C. T. 2000. "An Interview with Marvin Minsky," *Knowledge Management*, (June) pp. 26–28.
8. Zadeh, L. A. 1999. "From Computing with Numbers to Computing with Words—From Manipulation of Measurements to Manipulation of Perceptions," *IEEE Transactions on Circuits and Systems*, vol. 45, no. 1, pp. 105–119.
9. Rubin, S. H. 1999. "The Role of Computational Intelligence in the New Millennium," Plenary Speech, *Proceedings of the 3rd World Multiconference on Systemics, Cybernetics, and Informatics (SCI '99) and 5th International Conference on Information Systems Analysis and Synthesis (ISAS '99)*, pp. 3–13.
10. Rubin, S. H. 1998. "A Fuzzy Approach Towards Inferential Data Mining," *Computers and Industrial Engineering*, vol. 35, nos. 1-2, pp. 267–270.
11. Hindin, J. 1986. "Intelligent Tools Automate High-Level Language Programming," *Computer Design*, vol. 25, pp. 45–56.
12. Manning, B. 2000. "Smarter Knowledge Management Applications: LISP," *PC AI*, vol. 14, no. 4, pp. 28–31.
13. Gat, E. 2000. "LISP as an Alternative to Java," *Intelligence* (winter), pp. 21–24.

❖



Stuart H. Rubin

Ph.D. in Computer and Information Science, Lehigh University, 1988

Current Research: Intelligent systems; knowledge management.

Establishing a Data-Mining Environment for Wartime Event Prediction with an Object-Oriented Command and Control Database

Marion G. Ceruti
SSC San Diego

S. Joe McCarthy
Space and Naval Warfare Systems Command

ABSTRACT

This paper documents progress to date on a research project, the goal of which is wartime event prediction. The paper describes the operational concept, the data-mining environment, and the data-mining techniques that use Bayesian networks for classification. Key steps in the research plan are (1) implement machine learning, (2) test the trained networks, and (3) use the technique to support a battlefield commander by predicting enemy attacks. Data for training and testing the technique can be extracted from the object-oriented database that supports the Integrated Marine Multi-Agent Command and Control System (IMMACCS). The class structure in the IMMACCS data model is especially well suited to support attack classification.

INTRODUCTION

The ability to predict attacks and other hostile events during times of conflict is important to military commanders from the standpoint of readiness. The more advanced the notice and the more widespread the notification, the better able all echelons are to respond to threats efficiently and with the correct combination of forces.

The literature is replete with recent research results on data mining and data classification. (See, for example, [1, 2, 3, and 4].) Data mining, data classification, and data correlation are related to data fusion. As these techniques mature, better tools become available to model and to correlate data from complex operational scenarios. The purpose of this research is to create and extend a method to predict attacks on the U.S. Marine Corps using an object-oriented command and control database and data-mining techniques [5].

Data Mining

Data mining is the search for and extraction of hidden and useful patterns, structures, and trends in large, multidimensional, and heterogeneous data sets that were collected originally for another purpose. (See, for example, [4].) Data mining is an art that is supported by a considerable body of science, engineering, and technology. For example, data mining uses techniques from such diverse areas as data management, statistics, artificial intelligence, machine learning, pattern recognition, data visualization, and parallel and distributed computing. Data mining is possible today because of advances in these many fields; however, this multidiciplinary characteristic also makes data mining a difficult subject to teach and learn. Whereas the Structured Query Language (SQL) is inadequate to answer many complex queries, data mining can support searches for patterns in temporal and spatial databases in a more efficient manner. Data mining is important to the military because commanders and the analysts who support them cannot anticipate all future uses of information at the time of data collection.

Limitations of Data Mining

Whereas the goal of data mining is to identify hidden patterns, the search algorithms chosen for the particular task may miss an important and interesting pattern or even a class of similar patterns. A systematic method to preclude this problem is not available.

Similarly, there is no guarantee that any given data-mining effort will yield something new and useful, regardless of how many well-designed data-mining tools are used. This is because the data may not contain the desired patterns. Data mining is a search for observational data and the relationships between them, rather than the measurement of experimental data.

CONCEPT OF OPERATIONS

The concept of operations for a future system based on this research is (1) to use data-mining and data-classification algorithms to detect patterns associated with attacks (e.g., to identify factors that indicate an imminent attack) and (2) to correlate these patterns with current events with a view toward supplying military commanders with a prediction of the next attack and a confidence level that pertains to that prediction. A considerable amount of data associated with events that have preceded known attacks is required to model attacks, to search for common features, and to find these patterns in new data.

Success in this effort depends on a characterization of the circumstances that translate to well-defined observables that preceded past attacks. The more detailed the available knowledge, the better the resulting model, and the greater the probability that data instantiating critical variables can be collected. We expect that such detailed data for all variables will not be available prior to future attacks and that all available data may not be useful in predicting attacks (i.e., will function as "noise" in the analysis). Thus, the task involves identification of algorithms that can detect pre-attack features in clutter and the use of pattern recognition. Modern methods of statistical pattern recognition are sufficiently computationally oriented to use a larger dimensional space and are less sensitive to noise than older methods. Success in attack prediction will depend, at least in part, on how well these methods can be implemented with the available data.

GENERAL APPROACH

Hostile events can be characterized with respect to as many relevant variables as are deemed necessary and available to predict future attacks. An object-oriented message-traffic database can be analyzed for the occurrence of telltale signs of pending attacks. Our objective is to generate an event prediction (in terms of a probability) with a confidence value associated with it. Therefore, it is necessary to determine the combinations of events and observations that will have a higher probability of indicating a future attack. A baseline can be modeled from normal operational scenarios and from military events during times of conflict that do not constitute attacks per se.

The attack alarm-generation process and the reduction of false positives can be approached using constraints from models of known attacks. The identification of the appropriate features (and groups of features) that can flag imminent attacks is the most challenging part of the process. One approach is to explore the generation of a knowledge base encoded in Bayesian networks.

A literature search was conducted for publications on various subjects that relate to data mining, including algorithms and their applications. Data-mining algorithms can be used to identify complex patterns in the data that correlate well to hostile events. Criteria can be developed for sufficient correlation and confidence levels in data associations. For example, one metric that could be used is correlation strength, which is the ratio of the joint probability to the individual probability of observing a pattern [1].

BAYESIAN NETWORKS

Bayesian networks can be used to classify data into categories. Bayesian networks are:

- probabilistic networks,
- directed acyclic graphs that encode certain dependences between nodes that represent random variables,
- knowledge bases with knowledge in the network's structure and in its conditional probability table, and
- structures that can be used to infer causality.

Naive Bayesian Networks

A naive Bayesian network is a very simple structure in which all random variables representing observable data have a single, common parent node—the class variable. The naive Bayesian classifier has been used extensively for classification because of its simplicity, and because it embodies the strong independence assumption that, given the value of the class, the attributes are independent of each other.

Naive Bayesian networks work remarkably well considering that this independence assumption may not be valid from a logical standpoint. The performance of a naive Bayesian network can be improved with the addition of trees that provide augmenting edges to a naive Bayesian network by representing correlations between the attributes.

Tree Augmented Naive (TAN) Bayesian Classification Algorithm

SSC San Diego has access to SRI International's classifier algorithms developed under the Defense Advanced Research Projects Agency's High Performance Knowledge Base Program. For example, SRI's Tree Augmented Naive (TAN) Bayesian Classification Algorithm is a classifier algorithm based on Bayesian networks with the advantages of robustness and polynomial computational complexity [2 and 3].

Bayesian networks have some drawbacks that SRI has addressed in the TAN algorithm. In ordinary naive Bayesian networks, the variables (data) are assumed to be conditionally independent given the class. Logically, this is not always true. For example, suppose enemy troops are observed at location X and enemy tanks are observed at location Y. When using naive Bayesian networks, one assumes that these events are independent. However, both events may be part of the overall enemy battle plan. In the TAN algorithm, the trees provide edges that represent correlation between the variables.

Bayesian networks, especially with tree augmentation, are a suitable technology for data-mining classification and event prediction for the following reasons:

- First, one need not provide all joint probability values to specify a probability distribution for collections of independent variables [6].
- Second, one could mix modeling (e.g., explicit knowledge engineering for knowledge elicited from experts) with statistical data induction and

adaptivity. This mix would require fewer data values to induce better quality models.

- Third, one could use these models to compute the value of information. For example, having seen signs "A" and "B" of an imminent attack, what is the best information to collect next to confirm that hypothesis?
- Fourth, one could characterize explicitly the kinds of attacks. For example, given an attack of type "air attack," what are the most likely signals? These signals could be collected regularly to fill the database used as input into the TAN algorithm.

The TAN algorithm makes some tradeoffs between accuracy and computation. It approximates a probability distribution using some constraints on the complexity of the representation; however, it is extremely fast (low polynomial), efficient (one pass over the data), and robust (low-order statistics).

The TAN algorithm accepts data sets as input and induces Bayesian networks as output. Specifically, the TAN algorithm is intended to be used as a classification algorithm, which means that the input would be a file with tuples of the form $\{x_1, x_2, x_3, ..., x_n, c\}$ where the x_i, are values that variable X_i takes and c is the value that a class (C) variable can take. To set the range of each variable, the TAN algorithm needs an auxiliary file that contains a description of each variable, including the range of values representing the degree of intensity.

The TAN algorithm's output is a Bayesian network encoding of $P(C, X_n,...,X_1)$ in an efficient manner. To use TAN as a classifier, one simply computes $P(C|x'_n,....,x'_1)$. Given a new vector $X'_n,...,X'_1$ and having a probability distribution over c, one can select the event with highest probability as the one to classify. To compute the confidence in this value, the bootstrap method can be used [7].

The TAN algorithm outperforms naive Bayesian networks while maintaining its robustness and computational simplicity (polynomial vs. exponential complexity).

The TAN algorithm captures the best of both discrete and continuous attributes. Therefore, the TAN algorithm achieves classification performance that is at least as good as, and in some cases better than, models that use purely discrete or purely continuous variables. Studies at SRI have demonstrated that the TAN algorithm performs competitively with other state-of the-art methods.

TAN, and similar algorithms, can be made to perform the classification of certain battlefield situations for the Marine Corps. Much work needs to be done in this area, particularly with regard to data-set selection, data cleansing, and the refinement of the algorithm to meet specific needs.

In addition to the TAN algorithm, SRI has more general algorithms for inducing Bayesian networks that do not make the compromises that the TAN algorithm does. These algorithms try to fit the best distribution possible with no constraints. The disadvantage is that the computation of these models is slower; however, this may be acceptable and desirable in some cases. Algorithms can be implemented with the same data and the results compared.

GaussMeasurePredict Program

The GaussMeasurePredict program was developed by Nir Friedman to measure the performance of an induced TAN model. (See, for example, [2]).

The input of GaussMeasurePredict consists of the following items: (1) an induced Naive Bayesian network from TAN, (2) the name of the variable to predict, and (3) a test data set that contains instance information. When testing the Bayesian network model, the variable to predict is specified and known to be correct. Usually this will be the outcome of the class variable.

GaussMeasurePredict also has the option to calculate and display the probability of each class value for each instance in the input file. This feature is particularly useful for receiver operating characteristic (ROC) curves as well as for determining other statistics [8]. Thus, with this option, GaussMeasurePredict can output the probability distribution for each instance in addition to a summary.

The output of GaussMeasurePredict is a prediction of the accuracy of the network in the TAN Bayesian network .bn file. It can be used to predict the accuracy of other classifier algorithms as long as the output file matches the format of TAN's Bayesian network file.

GaussMeasurePredict is intended to be used to measure the accuracy of predictions and not to generate predictions for unlabeled instances. Unlike the TAN algorithm, GaussMeasurePredict does not accept instances with "?" for missing values in an instance input file. All variables must have filled values in each instance. However, because GaussMeasurePredict compares the induced Bayesian network to the test data set, it also can be used to infer the class of an unknown instance by filling in the class (Outcome) variable with a guessed value. Using the option described above, GaussMeasurePredict can output a predicted class probability for each class value. The class with the highest probability is the predicted class for that instance.

Fortunately, in the simplest case of attack predictions, only two values are possible for the class variable: ATTACK_LIKELY and ATTACK_NOT_LIKELY. In more detailed cases of attack predictions in which specific attack types are listed in the data-definition input file, the class variables may assume 2N values where N is the total number of attack types considered in the class. (The 2N arises from including the negation of the likelihood of an attack of each type.)

SOFTWARE IMPLEMENTATION AND PLANS

Data-mining software was tested for correct operation with clean data sets designed specifically for testing. The programs described below are included in the research environment. The software includes the TAN algorithm and the GaussMeasurePredict that uses the output of the TAN algorithm. Inputs to GaussMeasurePredict must be complete. Plans include the acquisition of additional algorithms that are designed to operate on incomplete data sets.

TAN 2.1 Availability

The TAN version 2.1 software and user's manual are available for download via file transfer protocol (FTP) from SRI's Web site: http://edi.erg.sri.com/tan/TANintro.htm. The user is required to register with a name and password. To obtain the TAN algorithm, Netscape is recommended and may be required. The Solaris CDE Web browser, HotJava, is not recommended to download TAN. The TAN user manual is included with the software (See, for example, [8]).

The TAN software was downloaded from SRI's Web site onto a Solaris SPARC Station 20 computer running the Solaris 2.7 UNIX operating system and using the Common Desktop Environment (CDE).

TAN 2.1 constitutes the main data-mining tool in the research environment of this project. TAN can be used as a base classifier and also as a method to fuse the output of other data-mining and classification algorithms. When algorithms have been tested and programmed, data visualization tools can be identified, tested, and used to view the data and to continue the pattern-recognition process.

GaussMeasurePredict Availability

The GaussMeasurePredict program is available along with the TAN software from SRI's Web site. The program is included with the TAN package and can be executed when files are "unzipped" and when the appropriate input files are available.

OBJECT-ORIENTED DATA IMPLEMENTATION

The object model, on which the Integrated Marine Multi-Agent Command and Control System (IMMACCS) database is based, is a detailed representation of the battlespace with objects derived from the March 1998 Urban Warrior Advanced Warfighting Exercise [9 and 10]. Object attributes and their associations, as well as class inheritance, are also described in [10]. The IMMACCS database uses the Unified Modeling Language symbolic representation method [10].

The IMMACCS database includes in its structure the following topics of interest to the Marine Corps: aircraft; ground vehicles; sea-surface vehicles; weapons and weapon systems; electronic devices of many kinds; terrain; bodies of water; logistics information; transportation infrastructure; various specialized units; personnel data; and most importantly for this application, military events. Class inheritance paths and allowed values are specified [10]. The use of an object-oriented database and the representation of military entities in object form provide a degree of interoperability and extensibility that allows multiple services to use and add to this common tactical picture [9].

The data sets for this data-mining effort will come from IMMACCS. The class structure in the IMMACCS data model is especially well-designed for adaptation to the attack/non-attack classification task. When data fill becomes available, especially for the attributes and object classes of interest, the IMMACCS database will be a very desirable data source for reasons described in the next subsection.

CONSTRUCTION OF TRAINING DATA SETS

The following discussion illustrates the strategy for constructing training data sets using certain IMMACCS object-oriented data classes as examples. The data-mining classification task is to identify the value of the Bayesian-network class variable of an unknown data set. Initially, two Bayesian-network class variables will be considered, "imminent attack likely" or "imminent attack not likely." To train the TAN algorithm, the value of the Bayesian-network class variable will be identified in the training data sets for both classes.

Various types of attacks and defenses are listed as allowed values (among others) in the MILITARY_EVENT object class in the IMMACCS database.

These are AIR_ATTACK, GROUND_ATTACK, AIR_DEFENSE, GROUND_DEFENSE, and SMALL_SCALE_ATTACK. Only instances that correspond to attacks from hostile forces on the Marine Corps will be considered. Any attack launched by the Marine Corps on hostile forces will not be counted in the "attack" category. In contrast, defenses by the Marine Corps against hostile attacks, whether the attacks are launched from the air or the ground, are likely to play a role in the overall model when they influence subsequent enemy attacks. For example, enemy commanders may select a battle plan that does not involve an air attack on an area with a strong Marine Corps air defense.

Several naive Bayesian networks can be induced, one for each attack type and one for the combined data for all attack types. For the combined attacks, the class variable can take multiple values, corresponding to the likelihood of a particular attack type and the likelihood that this attack type will NOT occur. Initially, all attack types will be assumed to be independent, although this is rarely true in actual battles. For example, ground attacks are more likely to follow air attacks at the same location than vice versa.

For the non-attack training instances, data associated with the other values of the MILITARY_EVENT object class will be used, such as WITHDRAWAL_EVENT, DELAYING_ACTION, AIR_REINFORCEMENT, or DRILL_EVENT. Other non-attack training instances also can be derived, for example, from the AIR_DEFENSE and GROUND_DEFENSE values, provided the instances pertain to events associated with enemy air defenses and ground defenses.

The date-time groups (DTGs) associated with each instance, both of attack and non-attack situations, will be noted and other data objects with the same DTGs (and with DTGs just prior to the event) will be included in the training data sets. The training data also could include objects present in the same vicinity as the attack or non-attack event that do not have DTGs. This will provide as comprehensive a description of the battlespace at the time and place of the attack as is possible, given the level of data granularity. This method of formulating training data sets can be extended by including in each data set the data that pertain to DTGs several days prior to the event to ascertain whether this will yield better results. The exact time span that each data set should cover is an open research issue.

Design Considerations in the Construction of Test Data Sets

Changes can be made in the test data sets, depending on the desired outcome of the test. For example, to determine how far in advance an attack can be predicted, the instances that pertain to an entire day immediately prior to the attack can be omitted systematically from test data sets. If the algorithm still makes the correct prediction, one can conclude, at least as far as that test data set is concerned, that an attack can be predicted 24 hours in advance. Similarly, if 2-days worth of data immediately preceding the attack can be omitted without a significant decline in the prediction accuracy, this is an indication that attacks can be predicted 48 hours in advance.

We expect, however, that omitting more and more data that pertain to the days just prior to an attack will cause the attack-prediction accuracy to degrade. The exact functionality of this degradation (linear, exponential, logarithmic, etc.) is another open research question. This type of testing can enable researchers to determine the number of days to include in the data collection and the specific data elements to be collected necessary to formulate as accurate a prediction as possible.

Test and training data sets will be formulated according to an n-fold cross-validation procedure. For example, to implement the first cycle of a five-fold cross validation with a data set consisting of 1,000 records, the first 800 records can be selected for training, with the last 200 records being reserved for testing. During the second phase of training and testing, the first 600 records and the last 200 records together will comprise the test data set, and the remaining records will be used for testing. In the third phase, the first and last 400 records will be used for training and the middle 200 for testing, etc. The advantage of this procedure is that it can be used to identify anomalies in the testing and training so that if the results are comparable for all five tests, a higher level of confidence in the method is obtained.

CONCLUSION

This paper describes a data-mining environment designed to support wartime event prediction using Bayesian networks to perform a data-classification task. The TAN algorithm was selected to induce a network using data extracted from an object-oriented database that contains information from exercise message traffic. Future work could include a user-friendly interface designed on top of the algorithms to provide automated input of selected data sets to the algorithm of choice. Success in this research project will pave the way for a more precise indication-and-warning system for the U.S. Marine Corps.

ACKNOWLEDGMENTS

The authors thank SSC San Diego's Science and Technology Initiative and the Defense Advanced Research Projects Agency for their financial support.

REFERENCES

1. Clifton, C. and R. Steinheiser. 1998. "Data Mining on Text," *Proceedings of the 22nd Annual IEEE International Computer Software and Applications Conference, COMPSAC'98*, pp. 630–635.
2. Friedman, N., D. Geiger, and M. Goldszmidt. 1997. "Bayesian Network Classifiers," *Machine Learning*, vol. 29, no. 2/3, November/December, pp. 131–163.
3. Friedman, N., M. Goldszmidt, and T. J. Lee. 1998. "Bayesian Network Classification with Continuous Attributes: Getting the Best of Both Discretization and Parametric Fitting," *Proceedings of the International Conference on Machine Learning '98*, ITAD-1632-MS-98-043.
4. Thuraisingham, B. M. 1999. *Data Mining: Technologies, Techniques, Tools and Trends*, CRC Press, Boca Raton, FL.
5. McCarthy, S. J. and M. G. Ceruti. 1999. "Advanced Data Fusion for Wartime Event Correlation and Prediction," *Proceedings of the 16th Annual AFCEA Federal Database Colloquium and Exposition*, AFCEA, pp. 243–249.
6. Charniak, E. 1991. "Bayesian Networks without Tears," *AI Magazine*, pp. 50–63.

Marion G. Ceruti

Ph.D. in Chemistry, University of California at Los Angeles, 1979

Current Research: Information systems analysis, including database and knowledge-base systems, artificial intelligence, data mining, cognitive reasoning, software scheduling and real-time systems; chemistry; acoustics.

S. Joe McCarthy

Ph.D. in Solid-State Electronics, University of Washington, 1973

Current Work: Assistant Program Manager for Processing and Analysis, Space and Naval Warfare Systems Command.

7. Friedman, N., M. Goldszmidt, and A. Wyner. 1999. "On the Application of the Bootstrap for Computing Confidence Measures on Features of Induced Bayesian Networks," *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics.*

8. Lee, T. J. and M. Goldszmidt. 1998. "TAN Tree Augmented Naive Bayesian Network Classifier Version 2.1 User Manual," http://edi.erg.sri.com/tan/TANintro.htm, pp. 1–27.

9. Alderson, S. L. 1999. "Urban Warrior Advanced Warfighting Experiment: Information Dominance in the Battlefield," *Proceedings of the 16th Annual AFCEA Federal Database Colloquium and Exposition*, AFCEA, pp. 213–228.

10. Leighton, R. and J. Pohl. 1998. The IMMACCS Object Model and Database, OBDATA00, November, IOM Version 1.5, Cal Poly, San Luis Obispo, CA.

❖

Thermal Pixel Array Characterization for Thermal Imager Test Set Applications

Ike Bendall, Ted Michno, Don Williams, Matthew Holck, and Richard Bates
SSC San Diego

José Manuel López-Alonso
Laboratorio de Termovision, Madrid, Spain

Robert J. Giannaris
Applied Technology Associates

Gordon Perkins and H. Ronald Martin
The Titan Corporation

ABSTRACT

An array of thermal emitters has been developed for use in a portable test set to enable field-testing of low-performance infrared imaging systems and seekers. It is not known if this technology can be used to evaluate the performance of state-of-the-art thermal imagers. This paper describes the preliminary measurements of thermal pixel array (TPA) performance. The radiant output of TPA was measured as a function of pattern size and drive voltage. Simple models were developed that agree with many aspects of the experimental data. Spatial and temporal noise characteristics of the TPA have been ascertained through three-dimensional noise analysis. Detection algorithms were used to compare images of test patterns produced by the TPA to images of similar test patterns produced by a standard blackbody.

INTRODUCTION

Infrared scene projection (IRSP) technology has advanced rapidly in the last few years in an effort to support testing of missiles and other munitions that use infrared seekers. Existing infrared scene generation technology is very expensive, with available scene generators falling in the million-dollar price range. These infrared scene projectors are prohibitively expensive for most infrared (IR) sensor test and evaluation applications. Low-cost alternative technologies would open the door to a much greater range of test applications.

A thermal imager test set using IRSP technology would have several advantages over traditional test sets consisting of a blackbody source and target wheel. Portable thermal imager test sets have a small number of target wheel positions. Test patterns must be installed to match sensor test requirements. IRSP technology eliminates the need for physical test patterns and allows the operator to generate test patterns appropriate for each sensor. Target wheels are generally too large to be effectively cooled. Blackbody sources can be controlled to maintain a constant temperature difference, but changes in the ambient temperature produce temperature changes that lie outside the camera's dynamic range. The IRSP arrays under investigation in this study are small and can be cooled with thermoelectric coolers. The use of IRSP technology in place of the blackbody/target wheel allows control of both the source and background temperatures and guarantees that the scene lies within the camera's dynamic range. The thermal imager testing community is developing improved methods of testing thermal imagers that do not use traditional test patterns. IRSP technology provides the tester with the flexibility to generate the test patterns appropriate to these alternative test procedures.

SSC San Diego has been funded through the Office of Naval Research to develop a low-cost thermal pixel array (TPA) for portable test set applications that provides a path to built-in test applications. The Real-Time Infrared (RTIR) TPA is a micro-electromechanical systems (MEMS) device consisting of a two-dimensional array of miniature IR heater elements (thermal pixels). In contrast to other IRSP technologies, the RTIR TPA is a silicon-based, micro-machined Complementary Metal Oxide Semiconductor (CMOS) array. This process yields a single chip device that is significantly less expensive than alternative approaches. Each IR heater is suspended over a micro-machined cavity and surrounded by pixel-specific electronics that allow rapid loading and retention of the image data. The micro-machined cavity thermally isolates the heater from the parent substrate, allowing each pixel to be individually set and maintained at a temperature different from that of its neighbors. Four heater elements are shown in Figure 1(A). Each heater element can be addressed independently of any other heater element. This allows the operator to vary both the shape and location of test patterns. This capability is shown in Figures 1(B), 1(C), and 1(D).

The RTIR TPA specifications were selected to meet dynamic scene requirements for missile testing and were not intended for use in a thermal imager test set. Minimum resolvable temperature difference (MRTD) is an important thermal imager figure of merit and is routinely measured during sensor evaluations. State-of-the-art thermal imagers have MRTDs of a few tens of milliKelvin at low spatial frequencies. Characterization of these imagers requires blackbodies with temperature resolutions that exceed those of the imager. Temperature resolutions of this scale exceed the RTIR TPA design specifications by at least an order of magnitude. In spite of the drawbacks of the RTIR TPA design, it was felt that it would be beneficial to compare the performance of this technology to a traditional thermal imager test set. This approach would provide insight into the feasibility of the RTIR TPA technology, help identify unknown problems, and provide a basis for developing thermal imager test set TPA performance specifications.

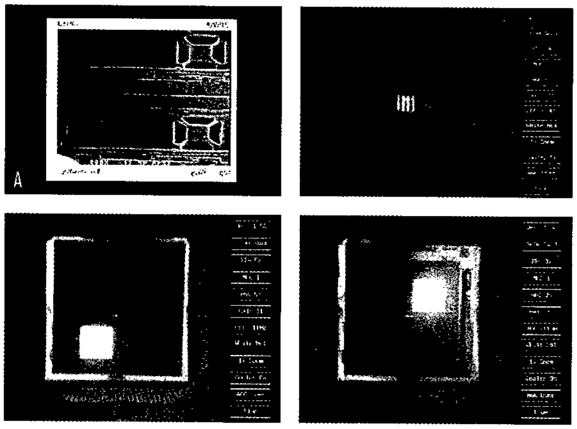


FIGURE 1. (A) Four micro-machined heater elements, (B) four-bar pattern in center of array, (C) square in lower left-hand corner, and (D) square moved to upper right-hand corner.

INSTRUMENTATION

The standard blackbody used in this comparison was furnished by Santa Barbara Infrared (SBIR). The telescope has a 6-inch aperture and a 30-inch focal length. Differential temperature resolution is ±3 milliKelvin when the unit is operated in the temperature difference mode. The thermal pixel array test set is shown in Figure 2. The RTIR TPA is a 128 by 128 array with pixel pitch of 88.6 microns.

The temperature range of the TPA is approximately 250°C with a thermal resolution of 0.250°C. The TPA area fill factor is 15%, and its emissivity is approximately 60%. The collimating telescope has a total transmission of 91% in the 3- to 5-micron band, a focal length of 233 mm, and a 50-mm aperture. The losses due to the fill-factor, emissivity, and telescope transmission result in an efficiency of 0.082 and an effective temperature resolution of 20 milliKelvin. A pixel non-uniformity correction capability is planned but is not currently available.

An Amber Galileo thermal imager with a 75-mm focal-length lens was used for these measurements. The Galileo is capable of extremely high frame rates; however, for this analysis, images were acquired at

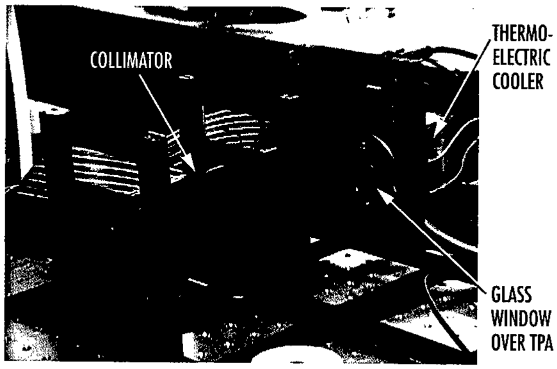


FIGURE 2. TPA blackbody.

the standard 30 Hz. The thermal imager was mounted on a rotation stage as shown in Figure 3. The thermal imager focus was adjusted to image the bar patterns from the SBIR blackbody. The imager was then rotated 90 degrees to view the TPA blackbody. The focus of the TPA bar patterns was achieved by adjusting the position of the TPA. This configuration allowed rapid collection of both TPA and SBIR images. Images were digitized with a Matrox Pulsar frame grabber with 8 bits of resolution.

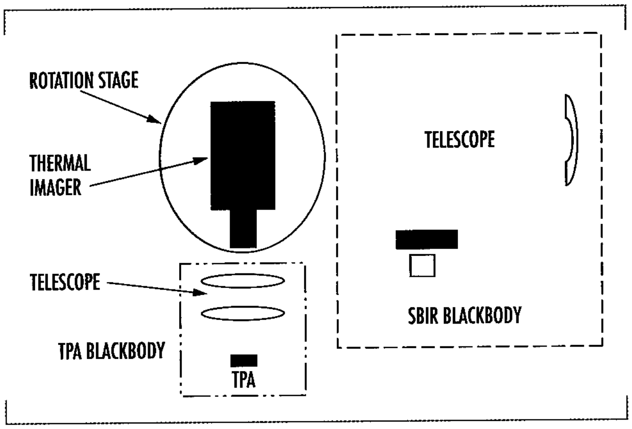


FIGURE 3. Diagram of experimental apparatus.

PATTERN SIZE AND VOLTAGE EFFECTS

Traditional test sets consist of a thermal source, a collimator, and a target wheel that holds the test patterns or masks. The wheel is physically separated from the blackbody source and its temperature is unaffected by changes in the temperature of the source. Changing the wheel's position does not affect the temperature difference between the blackbody and mask; therefore, temperature differences are independent of the pattern size. This is not necessarily true for a TPA blackbody. The thermal insulation provided by the micro-machined cavity does not completely isolate the heater from the parent substrate. Thermal conduction through the substrate affects the background temperature of the array and decreases the effective temperature difference (Figure 4). This effect may depend on both pattern size and control voltage.

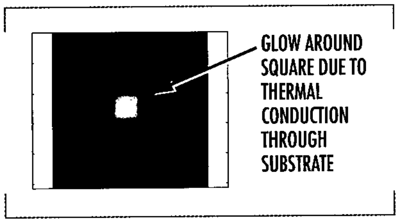

GLOW AROUND SQUARE DUE TO THERMAL CONDUCTION THROUGH SUBSTRATE

FIGURE 4. 40-pixel by 40-pixel heated area. It is apparent that the heat is not confined to the pattern area but is conducted into the surrounding area.

The first characterization task was to examine the relationship between pattern size and radiometric temperature. Three test patterns (two squares and a four-bar pattern) were selected for the analysis. The squares were generated by heating 30-pixel by 30-pixel and 6-pixel by 6-pixel regions on the array. The bar pattern consisted of four bars each 21 pixels long by 3 pixels wide. This pattern is consistent with the 7:1 aspect ratio of bar patterns used in MRTD measurements. Two measurements were

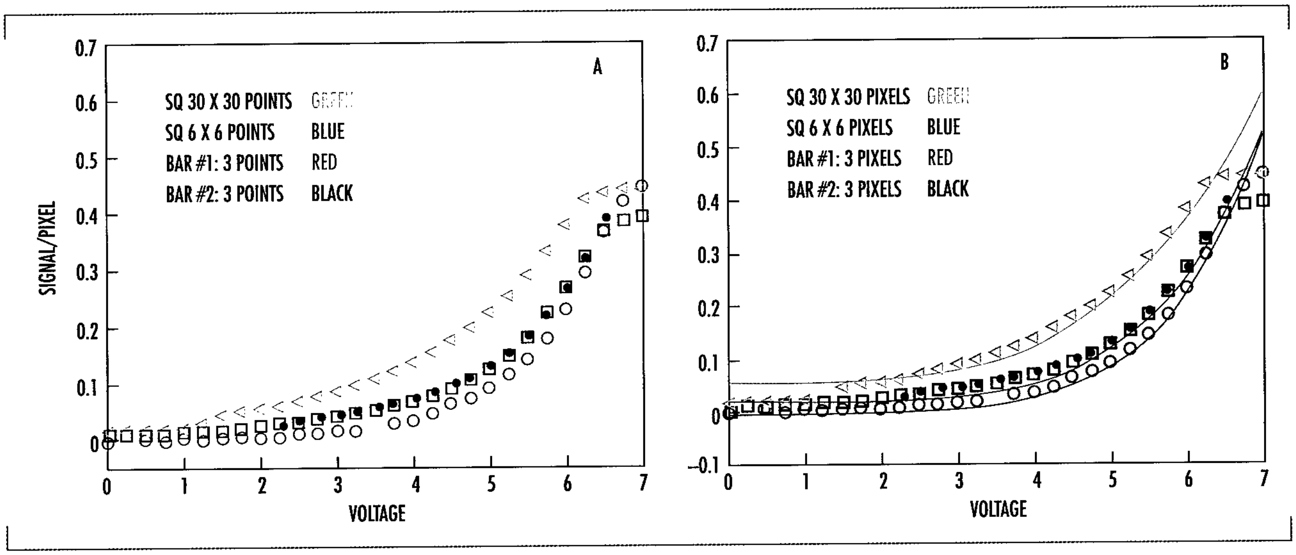


FIGURE 5. (A) Signal/pixel for three patterns (30 x 30 square, 6 x 6 square, and 21 x 3-bar pattern). Pattern size effects are evident. (B) Fit to data based on simplistic heating model. Model has three parameters and a nonlinear fit is used to achieve best fit. Good agreement is achieved between 2.5 and 6 V.

made with the bar pattern on separate days to evaluate TPA temporal stability. Images of each pattern were obtained as a function of voltage. The values of the pixels in each pattern were summed to produce a total signal. The signal was divided by the number of pixels comprising the pattern to produce a signal/pixel value. The signal/pixel values were compared for the three patterns. The results are plotted in Figure 5(A). A temperature dependence on pattern size is clearly evident. In contrast, the four curves would overlap if a traditional blackbody/target wheel test set had been used. It was encouraging that the two four-bar pattern curves (red and black) were in excellent agreement and that the shapes of the curves were similar for all three patterns. A simplistic model, relating the radiometric energy measured to voltage, was developed and used to fit the data. The model, which had three unknown parameters, was in excellent agreement with the data from 2.5 to 6 V (Figure 5(B)).

Thermal imagers suffer from blurring due to a reduction of the modulation transfer function with an increase in pattern spatial frequency. A traditional blackbody does not affect the pattern fidelity; therefore, any loss of fidelity can be attributed to the thermal imager. The blurring due to thermal conduction in the TPA test set results in a loss of pattern fidelity that must be separated from the degradation in image quality due to the thermal imager.

The investigation of the impact of thermal conduction on pattern fidelity was continued by examining the shape of square test patterns as a function of size and voltage. The results are summarized in Figure 6. Horizontal line profiles were taken through the center of the heated area. Line profiles of a 40 by 40 square as a function of voltage are plotted in Figure 6(A). Figure 6(B) compares line profiles for squares with sides of 10, 15, 20, 30, and 40 pixels at a constant 6 V. A parabolic curve, described by the equation below, was plotted through the peak of each curve.

$$S = S_{x_c} - a_{p,V}(V - V_T)(x - x_c)^2$$

where S is the pixel value, S_{x_c} is the peak pixel value, x_c is the pixel location at which the peak pixel value occurs, V_T is a threshold voltage (~3 V), and $a_{p,V}$ is a coefficient that can depend on pattern size and voltage. The curves shown in Figure 6 are generated by setting $a_{p,V}$ equal to a constant independent of pattern size or voltage. The curves appear to represent a reasonable fit to the data. This relatively simple relationship was unexpected and suggests that thermal conduction distortions can be readily understood, which is encouraging given the complexity of the TPA structure.

THERMAL MODELS

The results from the previous section suggested that a simple thermal conduction model might predict the effects of pattern size and control voltage on the array's temperature distributions. A finite-element analysis model was used to predict array temperature distributions. The TPA is a very complex structure, but for the first attempt, a simplistic model of the TPA was constructed. The TPA was assumed to be a homogeneous, isotropic material with a constant thermal conductivity and emissivity. It was further assumed that cooling occurs only through the bottom surface of the array and that the thermoelectric cooler maintains this surface at a fixed temperature. The objective of this analysis was to generate curves with trends similar to those shown in Figure 6. In particular, four features in Figure 6 were of interest: (1) the increase in peak temperature with pattern size, (2) the long tail in the unheated region, (3) the sharp transition between the heated area and the tail, and (4) the flat tops of the small squares. A typical result is shown in Figure 7.

The model results do show the increase in peak temperature with pattern size and the long tail in the unheated areas. The transition between the heated and unheated areas is not as sharp, and the tops of the small squares are more rounded than experimentally measured. A more complex model of the TPA is being constructed that should replicate these features.

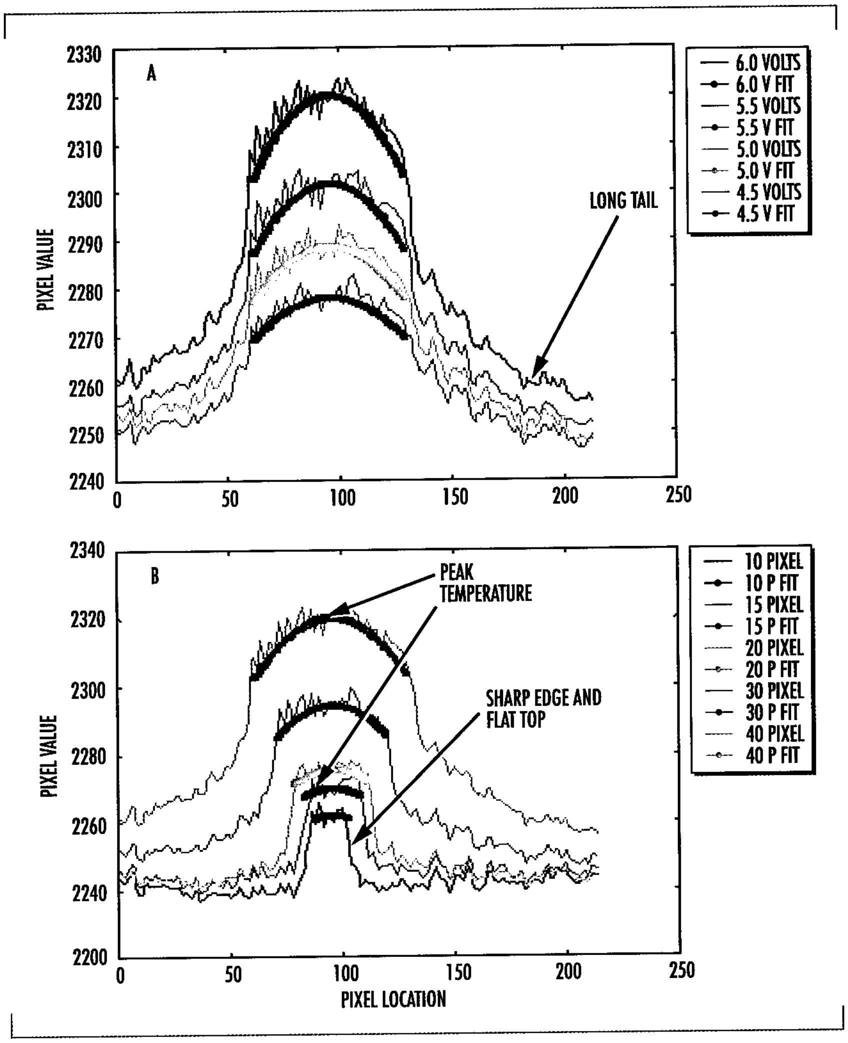


FIGURE 6. (A) Horizontal line profiles for a 40-pixel by 40-pixel pattern over a range of control voltages. (B) Horizontal line profiles at a fixed 6 V for square patterns of 10, 15, 20, 30, and 40 pixels. In both (A) and (B), the solid curve is parabolic fit.

NOISE BEHAVIOR

Noise is an important factor in thermal imager performance especially for tasks involving detection threshold measurements such as MRTD. The three-dimensional (3-D) noise model [1] provides an effective method of determining the noise characteristics. Image sequences of 30 frames were obtained from both the TPA and the reference blackbody. Thermal conduction through the TPA substrate distributes heat throughout the entire array. This low-frequency background is not apparent in the blackbody. For this reason, the low-frequency noise components were suppressed by means of a polynomial fit prior to the 3-D noise analysis. The results are summarized in Table 1.

The intrinsic noise of the blackbody should be small compared to that of the Galileo, and it is safe to attribute blackbody noise components in Table 1 to the Galileo. Inherent TPA noise is indicated by the increase between blackbody and TPA noise components. The magnitudes of the TPA and blackbody noise components are remarkably similar, with σ_{tvh} and σ_{vh} being the most significant noise components for both sources. This behavior is typical of staring thermal imagers, such as the Galileo.

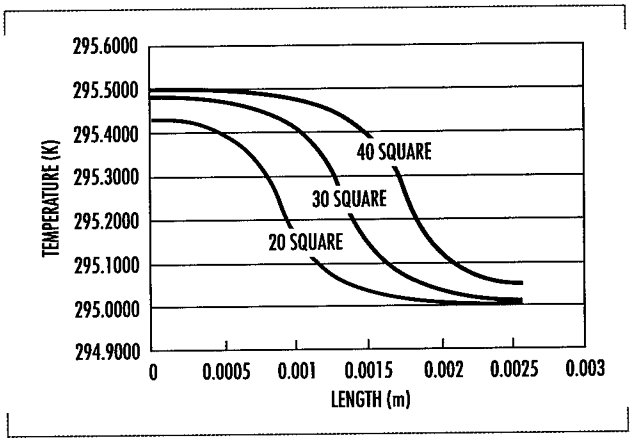


FIGURE 7. Finite-element analysis of TPA temperature profiles.

HUMAN VISION MODELS

In recent years, significant advances have occurred in the field of vision research to model the response of the human visual cortex. While human vision is far from solved, the principal mechanisms are understood. The visual cortex can be modeled as a collection of filters each sensitive to a restricted spatial frequency bandwidth. The model can be extended to compare filtered responses of two similar images and compute probabilities that a human observer will detect differences between the images. A visual cortex model developed by one of the authors [2] was used to compare high- and low-contrast bar patterns from a traditional blackbody and the TPA. A full description of the model and the analysis is beyond the scope of this paper; however, the model indicated that at low contrast the TPA and traditional blackbody images were indistinguishable to a human observer.

TABLE 1. 3-D noise analysis results.

	Blackbody (counts)	TPA (counts)
Sigma tvh	1.43	1.48
Sigma tv	0.23	0.24
Sigma th	0.16	0.22
Sigma vh	1.21	1.28
Sigma v	0.37	0.29
Sigma h	0.29	0.32
TOTAL	1.96	2.04

CONCLUSIONS

An assortment of measurements has been performed during the initial phase of the TPA characterization. In general, the results were extremely promising. Noise characteristics were in better agreement with a traditional blackbody than expected. Use of human vision models provided a novel characterization tool and indicated that TPA and blackbody images are very similar at low contrast. Crude estimates based on low-contrast images yield TPA MRTD measurements two to four times higher than MRTD measurements made with a traditional blackbody. This was better than expected, considering the poor temperature resolution of the RTIR TPA. Pattern blurring from thermal conduction is an important difference between TPA and traditional blackbodies. The effects of thermal conduction on pattern contrast must be understood or eliminated before a TPA-based thermal imager test set will be achievable. Simple thermal conduction models reproduce some of the experimentally measured features, but a more complete model is needed. Understanding the important factors affecting thermal conduction will help develop TPAs less susceptible to thermal distortions. Further investigation and development of the TPA is required, but the results are extremely promising and indicate that the TPA technology is a potential candidate for use in a thermal imager test set.

This technology may be the subject of one or more invention disclosures assignable to the U.S. Government, including N.C. #82901. Licensing inquiries may be directed to:
Harvey Fendelman
Office of Patent Counsel D0012
SSC SAN DIEGO
53510 Silvergate Avenue Room 103
San Diego, CA 92151–5765
(619) 553–3001

AUTHORS

Ted Michno
BA in Engineering Physics, Point Loma Nazarene University, 1996
Current Research: Microradiometry; infrared technology development.

Don Willliams
BS in Electrical Engineering, Northrup Institute of Technology, 1963
Current Research: MEMS for thermal infrared sources and for RF power measurement; passive millimeter-wave surveillance.

Matthew Holck
MS in Physics, San Diego State University, 1999
Current Research: Signal processing; image processing; computer automation.

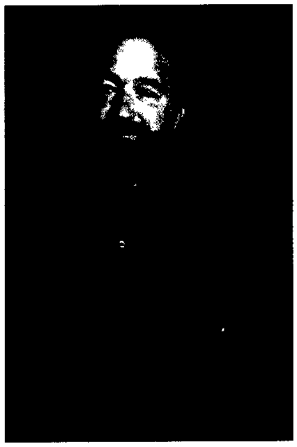
Richard Bates
BA in Physics, Loma Linda University, 1960
Current Research: Radiation-induced Irtran 2 absorption; polarization independent narrow channel (PINC) wavelength division multiplexing (WDM) fiber coupler fusing.

José Manuel López-Alonso
Graduate in Physics, Universidad Complutense de Madrid, 1994
Current Research: Characterization of thermal imagers, image quality evaluation.

Robert J. Giannaris
Ph.D. in Mechanical Engineering, Purdue University, 1972
Current Research: Infrared displays; microwave sensors; and hyperspectral sensors.

Gordon Perkins
BA in Physics, University of California at San Diego
Current Research: RF MEMS devices; millimeter-wave remote atmospheric sensing.

H. Ronald Marlin
BA in Physics, La Sierra University, 1959
Current Research: Microradiometry.



Ike Bendall
Ph.D. in Physics, Arizona State University, 1984
Current Research: Infrared and hyperspectral sensor characterization.

REFERENCES

1. D'Agostino, J. and C. Webb. "3-D Analysis Framework and Measurement Methodology for Imaging System Noise," *SPIE*, vol. 1488, pp. 110–120.
2. Alonso, J. and I. Bendall. 2000. "The Use of Vision Models for the Characterization of Scene Projector Devices," NATO TG12 Panel Group, (September), and private communication.

❖

Hyperspectral Imaging for Intelligence, Surveillance, and Reconnaissance

David Stein, Jon Schoonmaker, and Eric Coolbaugh
SSC San Diego

ABSTRACT

This paper highlights SSC San Diego contributions to the research and development of hyperspectral technology. SSC San Diego developed the real-time, onboard hyperspectral data processor for automated cueing of high-resolution imagery as part of the Adaptive Spectral Reconnaissance Program (ASRP), which demonstrated a practical solution to broad area search by leveraging hyperspectral phenomenology. SSC San Diego is now implementing the ASRP algorithm suite on parallel processors, using a portable, scalable architecture that will be remotely accessible. SSC San Diego performed the initial demonstrations that led to the Littoral Airborne Sensor Hyperspectral (LASH) program, which applies hyperspectral imaging to the problem of submarine detection in the littoral zone. Under the In-house Laboratory Independent Research (ILIR) program, SSC San Diego has developed new and enhanced methods for hyperspectral analysis and exploitation.

INTRODUCTION

The optical spectrum is generally considered to include the ultraviolet (200 to 400 nm), the visible (400 to 700 nm), the near infrared (700 to 1100 nm), and the short-wave infrared (1100 to 2500 nm). Sensors operating in these bands detect reflected light which is used to discriminate an object from its background and to classify it based on spectral characteristics. Spectral sensors capitalize on the color difference between objects and the background. A color video camera that divides the reflected light into red, green, and blue components is thus a simple spectral sensor. More complicated sensors break the spectrum into finer and finer bands and/or selectively tune to bands appropriate for a specific object or background. In general, a multispectral sensor, illustrated in Figure 1, is defined as a sensor using two to tens of bands, while a hyperspectral sensor, illustrated in Figure 2, is defined as a sensor using tens to hundreds of bands. Spectral sensors are divided into four types or approaches. Currently, the most common type is the "pushbroom" hyperspectral sensor. In this approach (Figure 2), a single line is imaged through a dispersing element so that the line is imaged in many different bands (colors) simultaneously. A second spatial dimension is realized through sensor motion. A second type is a multispectral filter wheel system in which a scene is imaged consecutively in multiple bands. A third type images multiple bands simultaneously using multiple chips (or multiple areas on the same chip). This approach uses multiple apertures or a splitting technique, such as a series of dichroic prisms or a tetrahedral mirror or lens. The fourth approach is the use of a Fourier transform spectrometer. The product of any of these sensors is an image cube as illustrated in Figure 3.

Hyperspectral Imaging at SSC San Diego

SSC San Diego has supported a number of hyperspectral programs over the last several years for a variety of government agencies, including the Defense Advanced Research Projects Agency (DARPA), the Spectral Information and Technology Assessment Center (SITAC), the Office of Naval Research (ONR), the Office of the Secretary of Defense (OSD), and the High Performance Computing Management Office (HPCMO). We have worked on DARPA's Adaptive Spectral Reconnaissance Program (ASRP), the goal of which was to demonstrate the detection of concealed terrestrial military targets and the cueing of a high-resolution imager. For ONR, we have been involved with maritime applications of hyperspectral sensors. Under OSD sponsorship, we have demonstrated the capabilities of hyperspectral remote sensing for search and rescue applications. For SITAC, we have provided ground truth measurements of ocean optical properties and illumination required for controlled experiments, and we have analyzed the bands required for optimal ocean imaging. The HPCMO is sponsoring our work to develop scalable and portable implementations of the ASRP algorithms. Under ONR and SSC San Diego In-house Laboratory Independent Research (ILIR) funding, we have developed new and enhanced methods for hyperspectral analysis and exploitation. Highlights of these efforts are described in more detail below.

Terrestrial Hyperspectral Remote Sensing

The DARPA ASRP successfully demonstrated the capability to detect military targets of interest in real time by using an airborne hyperspectral system to cue high-resolution images for ground analysis. SSC San Diego led all research, development, coding, and implementation of the end-to-end processing and critical hyperspectral detection and recognition algorithms. The algorithms and processing architecture developed are applicable to a broad scope of missions, targets of interest, and platform architectures. ASRP pushed the state of the art beyond simple detection of targets in the open, making detection of difficult, realistically positioned targets possible at low false alarm rates. Figure 4 shows the difficult environment, used by ASRP for real-time hyperspectral system demonstrations, that may be encountered during military operations. The variety of natural and man-made materials and the

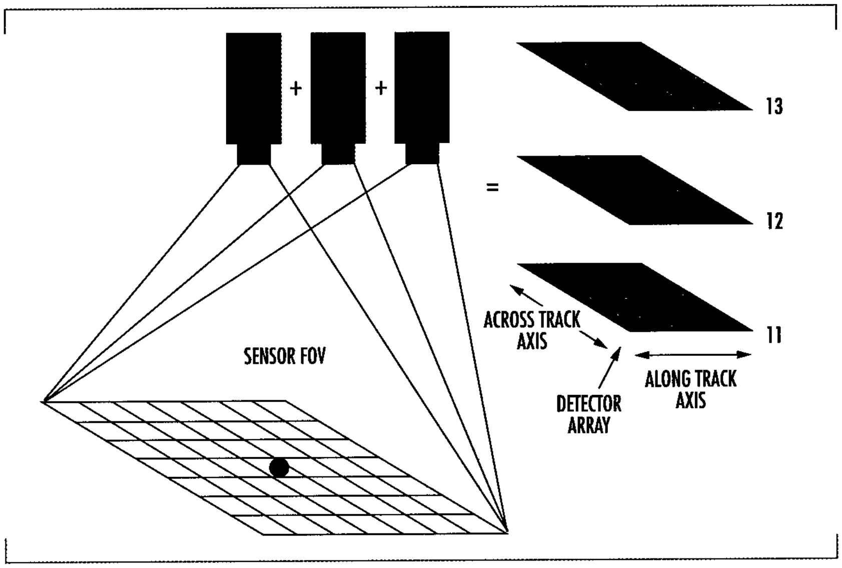


FIGURE 1. Schematic of three-band multispectral imaging camera.

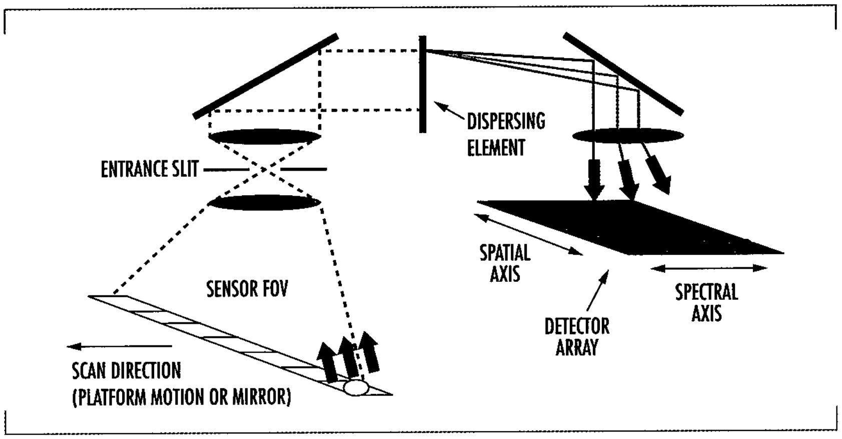


FIGURE 2. Schematic of a pushbroom dispersive hyperspectral sensor.

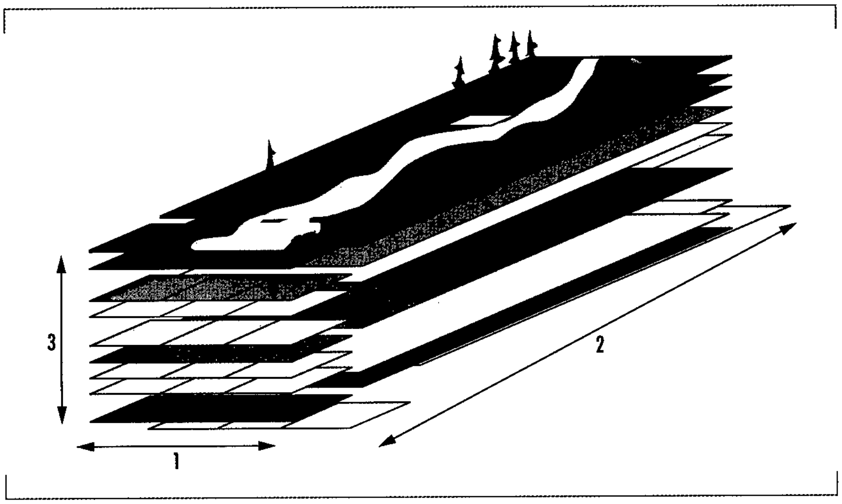


FIGURE 3. Hyperspectral image cube's cross-track, 1; along track, 2; and spectral dimension, 3.

variability of illumination combine to form a highly complex spectral detection challenge. Figure 5 compares the visibility of two targets in high-resolution imagery (top), in a red-green-blue (RGB) image (middle), and in the output of a detection statistic (bottom). These detections exemplify the ability of the hyperspectral system to identify target positions even when they may not be evident in traditional high-resolution imagery.

The High Performance Computing Management Office (HPCMO) has funded SSC San Diego, as part of the Hyperspectral Information Exploitation Project, to implement the ASRP hyperspectral algorithm suite and end-to-end processing on high-performance computer (HPC) platforms in a portable, scalable architecture accessible by a wide variety of Government users. Parallel processing capabilities will provide a new dimension for hyperspectral processing, allowing multiple hyperspectral algorithms to optimize target detection and recognition on massive data sets.

FIGURE 4. Three-color image of an ASRP hyperspectral image.

Maritime Sensor Systems

SSC San Diego has been instrumental in initiating and demonstrating the use of hyperspectral imagery for surveillance of the littoral. In 1996, SETS Technology, working with SSC San Diego, flew the SETS Technology Advanced Airborne Hyperspectral Imaging System (AAHIS) over submarines at the Pacific Missile Range Facility northwest of Kauai. The results of these flights led to the Littoral Airborne Sensor Hyperspectral (LASH) program.

LASH is an integrated optical sensor system that uses pushbroom scanning for the detection of submarines in the littoral environment. The LASH system consists of a

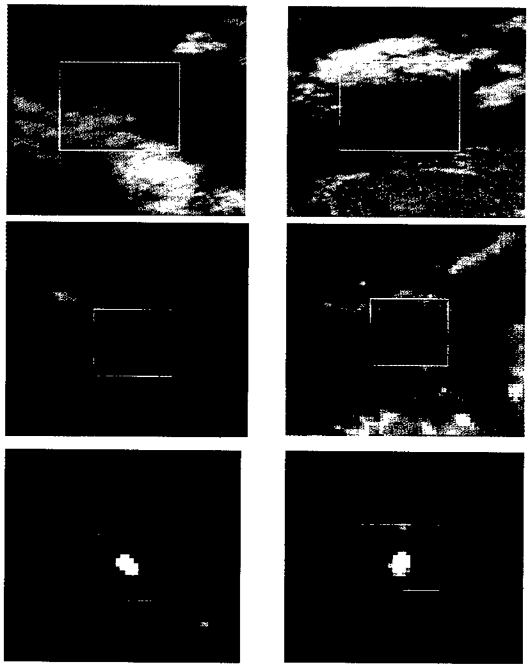


FIGURE 5. These figures show a high-resolution panchromatic imager (6-inch ground sample distance [GSD]) [top], and RGB image created from three hyperspectral bands (1-meter GSD) [middle], and one hyperspectral algorithm detection statistic image [bottom] for two different targets hidden along tree lines in shadow.

passive hyperspectral imager (HSI) assembly, an image processor, a data storage (archival) unit, a data display unit for operator use that incorporates the system monitoring, and control functions. The system is integrated into a modified ALE-43 (chaff cutter and dispenser pod) and mounted on a standard pylon at wing station 12 (Figure 6). All principal elements of the LASH system are contained within the pod. The units installed within the aircraft itself are limited to the system display processor, the power interface to the aircraft, the operator controls, and a global positioning system (GPS) antenna. This design was established to provide a system that could be considered independent of the individual aircraft tail number. It is estimated that all of the internal aircraft mounted units could be installed in less than 2 hours if necessary.

The passive and stabilized hyperspectral sensor collects both spatial (770 pixels) and spectral data (up to 288 pixels) on each instantaneous image increment. The data are binned by 2 spatially and 6 spectrally to give 385 spatial and 48 spectral channels. This imaged data is framed at 50 Hz, with each frame covering a 40-degree lateral field of view and approximately a 0.06-degree (1 milli-radian) field of view in the direction of flight. The data are simultaneously recorded in the archival storage system, processed by the image processor, and presented in a pseudo-color waterfall display to the operator. The processing system evaluates the data sensed in near real time using both spectral and spatial processing, and it provides a "frozen" display of the target along with its position in longitude and latitude. A stabilization system automatically adjusts the sensor so that it compensates for aircraft roll, pitch, and yaw. A "point to track" option forces the stabilization system to point the sensor along a predetermined track (otherwise the sensor points directly down).

FIGURE 6. LASH pushbroom hyperspectral imager mounted on the wing of a P3 aircraft.

These sensors can perform a wide range of ocean sensing tasks. Targets range from submarines and sea mines for military applications, to chlorophyll and sediment load in physical oceanographic applications, to schools of dolphins and whales in marine biology applications. Figure 7 demonstrates the ability of the sensor to image a pod of humpback whales. In these applications and others, a common goal is to detect an extremely low-contrast target in a high-clutter background.

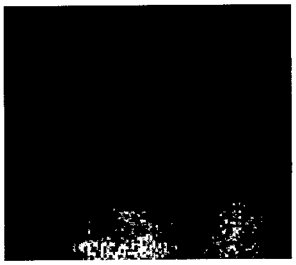


FIGURE 7. A pod of humpback whales imaged using the AAHIS sensor, a precursor to LASH.

Ocean Environmental Measurements

Hyperspectral systems such as LASH are being developed that use spectral and spatial processing algorithms to discern objects and organisms below the sea surface. The performance of such systems depends on environmental and optical properties of the sea. An instrument suite, the Portable Profiling Oceanographic Instrument System (PorPOIS), was developed to ascertain and quantify these environmental and hydro-optic conditions. Profiling of the downwelling irradiance leads to a value of the diffuse attenuation coefficient, k_d, for the water column. Measurements of the beam absorption, a, and attenuation, c, provide information about the non-pure water absorption and scattering characteristics of the water. Measurement of the backscatter at different wavelengths determines what fraction of the downwelling photons is scattered back toward space. These and a number of other measurements made by PorPOIS allow for a thorough characterization of the water body. These data are used in the LASH program to optimize parameters of the processing algorithms and to predict the performance of the sensor by using modeling software that requires these oceanographic data as inputs.

The PorPOIS system is deployed on two submersible cages and a surface data-gathering station. The instruments are controlled and the data collected on a laptop computer running a Windows-based control and data acquisition software package, the Sensor Interface Display (SID), developed at SSC San Diego. The instruments (Figure 8) used to measure surface conditions and ship location include a wind transducer (anemometer), a magnetic compass, a surface irradiometer, and a GPS receiver. There are currently seven instruments used to measure optical and environmental conditions below the sea surface. These instruments include a downwelling and upwelling irradiometer (Biospherical Instruments PER600 and PER700), an upwelling radiometer (PER600), a transmissometer (Seatech), an absorption and attenuation meter (WETLabs ac-9), a conductivity-temperature-depth (CTD) (SeaBird Electronics SBE-19), a fluorometer (WETStar), and a backscattering meter (HobiScat-6). The devices are bundled in a single beehive-type stainless-steel profiling cage as shown in Figure 9. The cage is suspended from a davit via the underwater cable. The SeaBird SBE-32 carousel water sampler (Figure 10) holds twelve 2.5-liter bottles and the SBE-19 CTD. It uses the same underwater cable as the profiling cage. Deployment of the cage is nearly identical to that of the instrument cage. A deck unit mounted in the control rack translates the CTD information from the carousel and transfers the data to SID. This allows the user to capture water samples from target depths by monitoring the position of the carousel as it travels through the water column. New instruments can be added to the configuration as required.

Sample PorPOIS products are shown in Figures 11 and 12. Figure 11 shows downwelling irradiance at 490 nm measured off San Clemente Island, CA. These data are used to determine the rate of attenuation of irradiance at 490 nm, k_{490}, as shown in Figure 12. Optical depth, $1/k_{\lambda}$, is defined as the depth at which surface irradiance of wavelength λ diminishes by 1/e. System performance is parameterized in terms of optical depth.

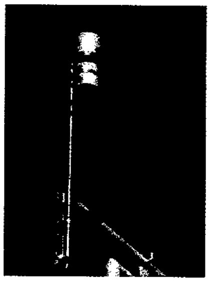


FIGURE 8. The Biospherical Instruments PRR-610 surface irradiometer, the NEXUS wind transducer, and the NEXUS magnetic compass are used to measure surface conditions.

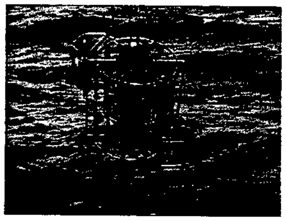


FIGURE 9. Submersible cage containing instruments used to measure ocean optical properties.

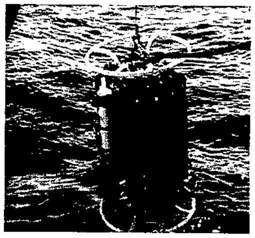


FIGURE 10. Submersible cage containing a CTD and water collection bottles used to measure absorption and scattering as a function of depth.

SSC San Diego ILIR and ONR-sponsored Research on Hyperspectral Algorithms

Pre-processing transforms are a common initial step in the processing of hyperspectral imagery that is performed in order to determine spectra of the fundamental constituents of the scene or for data compression. The principal component transform is based on minimizing loss in mean-square error, and the vector quantization (VQ) transform is based on minimizing the worst-case angle error between a datum and its projection onto a subspace. These transforms may have unintended consequences on the signal-to-noise ratio (SNR) of a target of interest. We have evaluated the loss in SNR that may result from applying a linear transform and developed several new transforms that use different knowledge of the signals of interest to reduce the loss in SNR in comparison with commonly applied transforms. Figures 13 and 14 illustrate the detectability of an underwater target in data that has been transformed using vector quantization and one of the newly defined transforms, whitened vector quantization (WVQ), that uses no signal information. Clearly, the WVQ algorithm can reduce the dimension of the data and preserve the target SNR for these

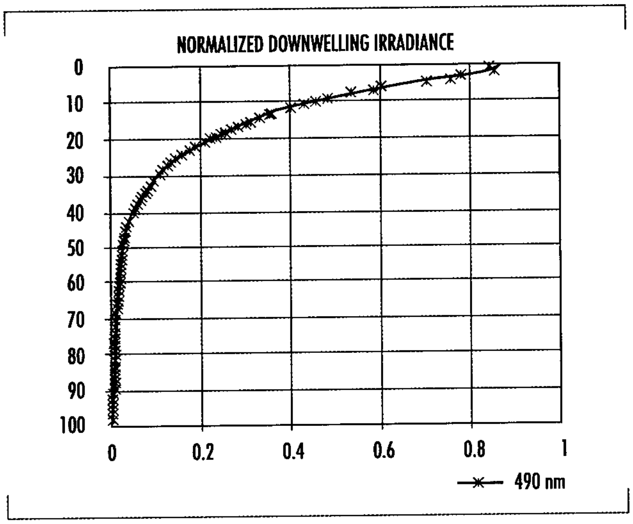


FIGURE 11. Plot of downwelling irradiance at 490 nm as a function of depth as measured using PorPOIS in waters off San Clemente Island, CA.

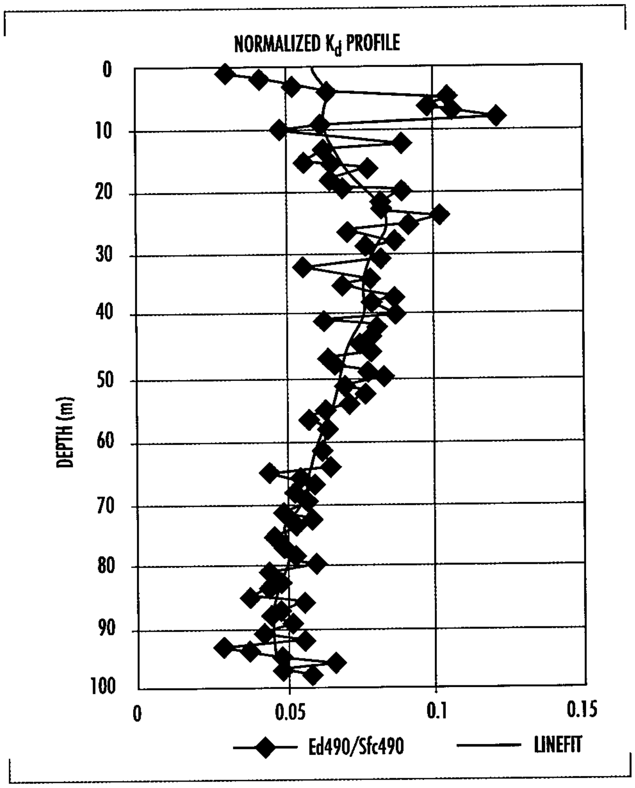


FIGURE 12. Rate of attenuation of downwelling irradiance at 490 nm derived from PorPOIS measurements of downwelling irradiance.

data. The transformed data are processed here with the Reed-Xioali (RX) quadratic anomaly detector. The enhanced discrimination of the target at lower dimension using the WVQ algorithm arises from the fact that the performance of quadratic detectors improves for a given SNR if the dimension is reduced.

Linear unmixing and image segmentation are common means of analyzing hyperspectral imagery. Linear unmixing models the observed spectra as

$$y^{ij} = \sum_{k=1}^{d} a_k^{ij} e_k \text{, such that} \sum_{k=1}^{d} a_k^{ij} \leq 1 \text{ and } 0 \leq a_k^{ij} \leq 1.$$

The spectral vectors, e_k are known as endmembers, and a_k^{ij} is the abundance of the k^th material at pixel (i,j). There are several means available for estimating the endmembers. The abundances are usually estimated by solving the constrained least-squares problem.

Image segmentation typically models the observation vector as arising from one of several classes, such that each class has a multi-variate normal distribution. The number of classes, d, is selected and the mean and covariance of the classes $\{(\mu_{\ell k}, \Sigma_k) \mid 1 \leq k \leq d\}$ are estimated from the hyperspectral data. The expectation maximization and the stochastic expectation maximization algorithm are two methods of estimating these parameters. Given the parameters and the probability of each class, the data may be classified by assigning y^{ij} to the class that, conditioned on the observation, is most likely. This computation is carried out using Bayes Law.

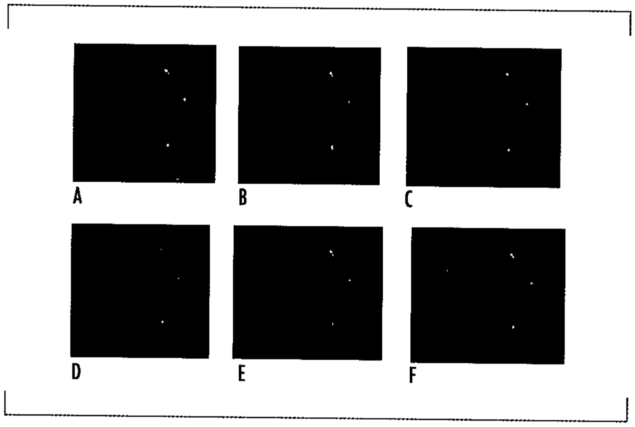


FIGURE 13. The RX algorithm applied to VQ-transformed 48-band hyperspectral imagery transformed to 48, 36, 20, 12, 9, and 7 dimensions (A through F, respectively).

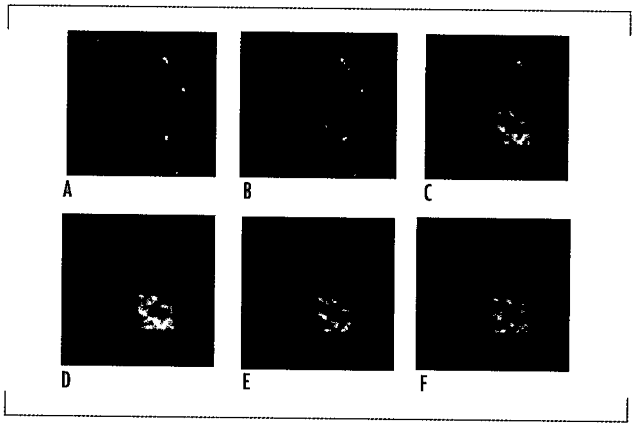


FIGURE 14. The RX algorithm applied to WVQ-transformed 48-band hyperspectral imagery transformed to 48, 36, 24, 8, 4, and 2 dimensions (A through F, respectively).

We have generalized the linear unmixing and image segmentation approaches in the development of the stochastic compositional model. We assume an A×B image of multivariate date: $y^{ij} \in R^n$, $1 \le i \le A$, $1 \le j \le B$. The stochastic compositional approach models each observation vector as a constrained linear combination of normally distributed random variables. Let d be the number of classes, and let $N(\mu_k, \Sigma_k)$, $1 \le k \le d$ denote the normal distribution with mean μ_k and covariance Σ_k then

$$y^{ij} = \sum_{k=1}^{d} a_k^{ij} x_k^{ij} \text{ such that } x_k^{ij} \sim N(\mu_k, \Sigma_k),\ 0 \le a_k^{ij} \le 1, \text{ and } \sum_{k=1}^{d} a_k^{ij} = 1. \quad (1)$$

To account for scalar variation in the illumination, we also consider the model that uses an inequality constraint:

$$y^{ij} = \sum_{k=1}^{d} a_k^{ij} x_k^{ij} \text{ such that } x_k^{ij} \sim N(\mu_k, \Sigma_k),\ 0 \le a_k^{ij} \le 1, \text{ and } \sum_{k=1}^{d} a_k^{ij} \le 1. \quad (2)$$

For given parameters (μ_k, Σ_k), $1 \le k \le d$, and given abundances $\alpha = (a_1, \cdots, a_d)$, let (dropping the pixel indices) $\mu(\alpha) = \sum_{k=1}^{d} a_k \mu_k$, and $\Sigma(\alpha) = \sum_{k=1}^{d} a_k{}^2 \Sigma_k$. Then, $y^{ij} \sim N(\mu(\alpha), \Sigma(\alpha))$. Maximum likelihood abundance estimates are thus obtained by solving

$$\hat{\alpha}^{ij} = \arg(\max(\frac{1}{|\Sigma(\alpha)|^{0.5}\,(2\pi)^{n/2}} \exp\left(\frac{-1}{2}\left(y^{ij} - \mu(\alpha)\right)\Sigma(\alpha)^{-1}\left(y^{ij} - \mu(\alpha)\right)\right). \quad (3)$$

Let $X = (x_1,\ldots,x_d)$; the maximum likelihood estimates of the decomposition of the observation into contributions, x_k from the classes is obtained by solving

$$\hat{X} = \arg(\max(p(X \mid y,\alpha,\mu_k,\Sigma_k)))$$

$$= \arg\left(\max\left(\prod_{k=1}^{d} \frac{1}{(2\pi)^{n/2}\,|\Sigma_k|^{1/2}} \exp\left(\frac{-1}{2}(x_k - \mu_k)'\Sigma_k^{-1}(x_k - \mu_k)\right)\right)\right)$$

$$\text{such that } y = \sum_{k=1}^{d} a_k x_k\,. \tag{4}$$

The stochastic compositional model and deterministic linear unmixing have been compared by using simulated hyperspectral imagery. Class statistics were estimated from hyperspectral imagery by using the stochastic expectation maximization algorithm. Using these parameters, a set of simulated hyperspectral imagery was generated so that the mixing proportions of the classes were known. The test data were then unmixed by using both deterministic unmixing (with the class means as endmembers) and by stochastic compositional modeling, such that the class parameters were estimated using the expectation maximization algorithm. Figure 15 compares the error in the abundance estimates of one of the classes using the two methods. In this example, the stochastic compositional model reduces the abundance estimation error by a factor of two to three. Work is ongoing to compare the performance of detection algorithms emanating from the segmentation, linear unmixing, and stochastic compositional models.

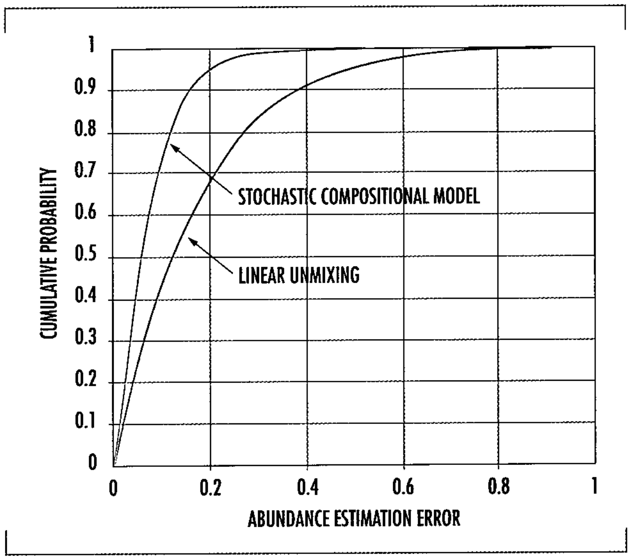
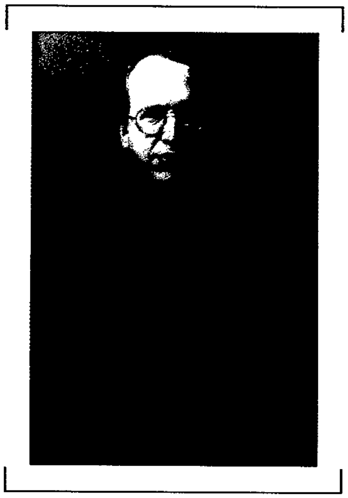


FIGURE 15. A comparison of the absolute error in the abundance estimate using linear unmixing and stochastic compositional modeling.

SUMMARY

SSC San Diego has been involved in many aspects of hyperspectral imaging. We are making important contributions in the areas of real-time processing implementations, system design for a variety of missions, environmental characterization, and the development of new models and methods. SSC San Diego is continuing to work across the Department of Defense (DoD)/Intelligence communities to bring mature hyperspectral technologies to the warfighter, making this unique source of critical information more widely available and user friendly.



David Stein

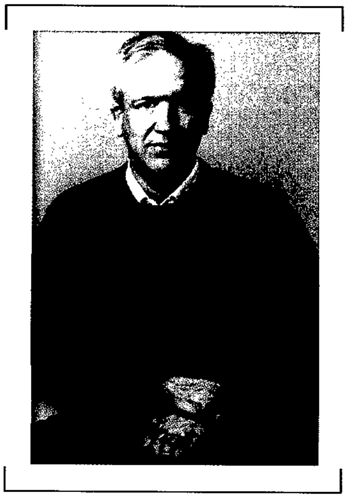
Ph.D. in Mathematics, Brandeis University, 1986

Current Research: Multidimensional statistics; detection theory; hyperspectral algorithms; remote sensing of littoral processes.

Jon Schoonmaker

BS in Physics/Mathematics, University of Oregon, 1985

Current Research: Hyperspectral systems; data analysis and algorithm development; remote sensing of littoral processes.



Eric Coolbaugh

MS in Oceanography and Meteorology, Naval Postgraduate School, 1989

Current Research: Hyperspectral imaging systems; high-performance computing; hyperspectral algorithms.

Surface Plasmon Tunable Filter for Multiband Spectral Imaging

**Stephen D. Russell, Randy L. Shimabukuro,
Ayax D. Ramirez, and Michael G. Lovern**
SSC San Diego

Yu Wang
Jet Propulsion Laboratory

ABSTRACT

The SSC San Diego Advanced Technology Branch and the Jet Propulsion Laboratory have been developing a novel technology that can be applied to multiband imaging. The surface plasmon tunable filter (SPTF) uses color-selective absorption by a surface plasmon at a metal-dielectric interface to achieve its optical selectivity. If an electro-optic material is used as the dielectric and a voltage is applied to change the surface plasmon resonance, the reflected light can be modulated, i.e., the photons at surface plasmon resonance will be absorbed and the photons out of the resonance will be totally reflected. Therefore, the applied voltage controls the reflection spectrum, and an electrically tunable color filter is formed. This paper details progress in developing SPTF technology as a replacement for discrete filters. This technology will allow multi-band or hyperspectral imaging with a single filter/camera system.

BACKGROUND

An important aspect of theater missile defense is the multiband spectral characterization of plume radiation during the boost phase of a missile. Current Ballistic Missile Defense Organization (BMDO) plans call for study of the utility of a dual-mode ultraviolet (UV) and mid-wave infrared (MWIR) seeker. Combining the conventional MWIR sensor with shorter wavelengths provides increased information content for the image and can aid in optical target characterization. However, even dual-mode seekers have potential problems. Onboard optical seekers are subject to some vehicle self-interference. Sources of optical interference include outgassing of vehicle contaminants, and by-products of the vehicle plume and attitude control systems, especially if solid aluminized propellants are used. Carbon particles are commonly present in the exhaust plume of kerosene liquid-oxygen (LOX) motors used by Atlas-type rockets. Once formed, carbon may contribute a continuum-like feature to the optical radiation of a rocket exhaust plume, especially in the near-UV [1]. A carbon monoxide–oxygen chemiluminescence mechanism may also be a source of radiation for the Atlas propellant because carbon dioxide is a large plume exhaust species and atomic oxygen is formed in the shear layer of the plume where the ambient oxygen molecules are dissociated [2]. Such optical interference effects lead to an increased background radiation level for the seeker in all spectral bands, but are most problematical in the infrared. Sensor confusion may also be caused by deliberate countermeasures. Therefore, multi-spectral imaging is important for ground-based imagery for optical signature characterization and onboard seekers.

One approach for multi-spectral imaging uses an imaging spectrometer that acquires images in many contiguous spectral bands simultaneously over a given spectral range. By adding wavelength to the image as a third dimension, the spectrum of any pixel in the scene can be calculated. These images can be used to obtain the spectrum for each image pixel, which can identify components in the target. The most common method of image spectroscopy is changing fixed dichroic filters. Existing systems suffer from large size and weight and operate slowly (approximately a millisecond). Several tunable filters have been proposed, but they all have severe problems. For example, the acousto-optic tunable filter is power-hungry (in kilowatts), while the liquid crystal tunable filter is slow (approximately tens of milliseconds for nematic liquid crystals) and has low efficiency.

The Advanced Technology Branch at SSC San Diego and the Jet Propulsion Laboratory (JPL) have been developing a novel technology that can be applied to BMDO's needs for multi-spectral imaging. The surface plasmon tunable filter (SPTF) described in this paper uses color-selective absorption by a surface plasmon at a metal-dielectric interface to achieve its optical selectivity. If an electro-optic (EO) material is used, an applied voltage can control the resonant frequency of the surface plasmon, and an electrically tunable color filter is formed [3, 4, 5, and 6]. The technology may replace discrete filters and allow for multi-spectral or hyperspectral imaging with a single filter/camera system. This feature is particularly important if minimal payload weight and volume is desired for imager or seeker systems on rockets or missiles.

SURFACE PLASMON TUNABLE FILTER

The surface plasmon (SP) has been studied since the 1960s. It is a collective oscillation in electron density at the interface of a metal and a dielectric [7]. At SP resonance, the reflected light vanishes. This resonance is attenuated total reflection and depends on the dielectric constants of the metal and the dielectric. If an EO material is used as the dielectric and a voltage is applied to change the SP resonance condition, the reflected light can be modulated [8 and 9]. Using this principle, an SP spatial laser light modulator with a contrast ratio greater than 100 has been reported [10]. If we consider the SP light modulator in frequency space, the photons at the SP resonance frequency will be absorbed by the free electrons in the metal, and the photons away from the SP resonance will be totally reflected. If a voltage is applied to the EO material, the resonance frequency will change, and a tunable filter is formed. The SP tunable notch filter was invented based on this voltage-induced color-selective absorption [11 and 12]. Figure 1 schematically shows a reflective-mode SPTF.

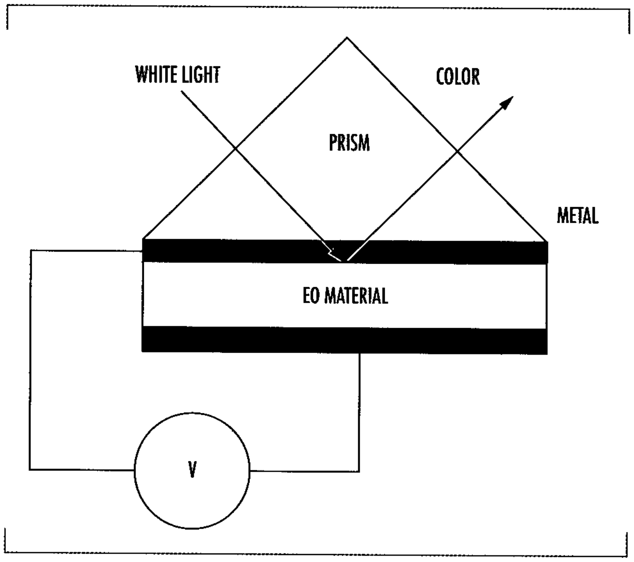


FIGURE 1. Reflective SPTF.

The structure of the SPTF in Figure 1 shows white light incident on the metal-EO interface using a high-index prism (SF57 glass) for coupling. The color of the reflected light is determined by the SP resonance that is a function of the dielectric properties of the materials. Using a thin (55-nm) layer of silver and a liquid crystal (Merck E49) as the EO material, a narrowband SP resonance is obtained (Figure 2). Note that as the applied voltage is increased from 0 to 30 V, the SP absorption shifts from red to violet.

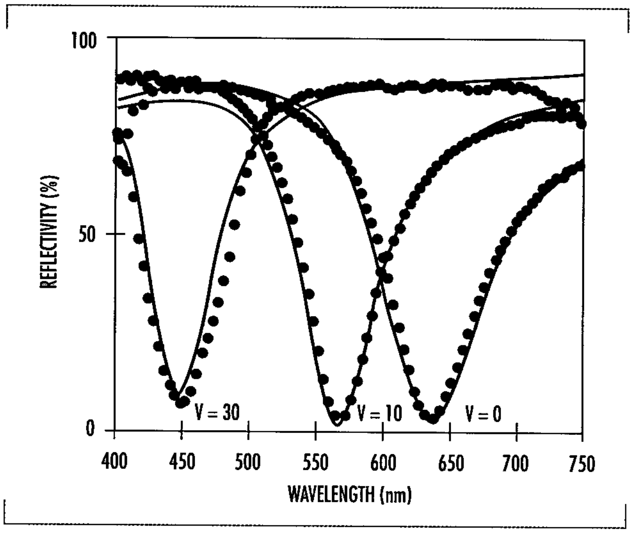


FIGURE 2. SPTF reflection spectra.

Figure 3 shows a symmetric geometry of metal/EO/metal used to form a transmissive filter. Two high-index glass prisms are used for the coupling with a thin metal film evaporated on each prism, and an EO material sandwiched between the two prisms. The thickness of EO material layer is less than 1 wavelength. When an SP wave is excited on one side of the metal/EO material interface by the incident photons, the energy of the resonance photons convert into the motion of free electrons of the metal film. The optical field penetrates the thin EO layer and excites another SP wave with the same frequency at the other EO/metal interface because of the symmetric structure. The resonance photons will then re-radiate out as transmitted light. When a voltage is applied to the EO material, the index of the EO material changes, leading to a change of the SP resonance frequency and the transmission spectrum. Theoretical calculation shows that for two silver films separated by a 150-nm EO material layer (Merck E49), a change in the index of the EO layer from 1.5 to 2.0 leads to transmission peak shifts from 450 to 650 nm.

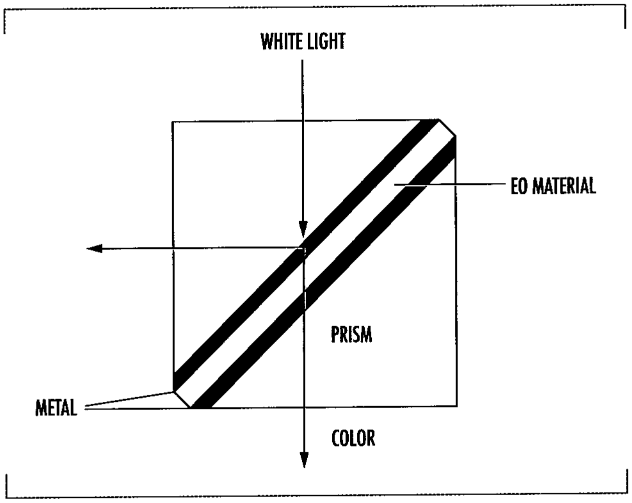


FIGURE 3. Transmissive SPTF.

Varying the thickness of the dielectric layer between the two metal films can also change the coupling mechanics. Using a symmetric geometry similar to what was used in Figure 3, a SPTF can be constructed using a changeable air gap to select the spectrum. Figure 4 shows the theoretical calculation of reflectivity vs. wavelength of the Air Gap SPTF and its effective tuning ability. Using silver as the metal films, when the thickness of the air gap changes from 300 to 5000 nm, the peak transmission shifts from 400 to 1600 nm. Though the structure of the Air Gap SPTF is schematically similar to the Fabry-Perot filter, the physics is totally different. The photons are incident at an angle greater than the critical angle, and two metal films must be used to generate the SP resonance. Furthermore, the tunable range runs from 400 to 1600 nm and is not limited by 2X as the Fabry-Perot filter requires. The SPTF can also be configured to operate based on angle of incidence [13].

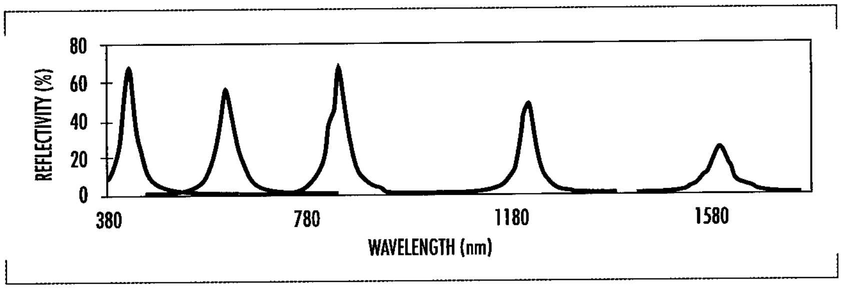


FIGURE 4. Tuning of the Air Gap SPTF.

FUTURE ADVANCES

A major advantage of SPFT technology is the ability to integrate it with various optical sensors and detectors. These products include state-of-the-art miniature photo-multiplier tubes available commercially (e.g., Hamamatsu R5600), microelectronic photo-multipliers [14 and 15], and solid-state detectors such as charge-coupled devices (CCDs) and active pixel sensors [16]. Compared with an acoustic-optic tunable filter and liquid crystal tunable filter, the SPTF is lightweight, low-power, and works in a wide temperature range. If the glass material is chosen so that its thermal expansion matches the thermal expansion of the EO material, this device works in a wide temperature range (-200 to +200°C. Though liquid crystal material was used in these experiments, the liquid crystal material can be replaced by solid-state EO materials such as potassium di-hydrogen phosphate (KDP), potassium titanyl phosphate (KTP), ethylene oxide (EO) polymers, organic crystals, and organic salts. If a solid-state material is used, the SP modulator can reach very fast (less than 1-μs) modulation speeds. Materials optimized for near-infrared (IR) and mid-IR can also optimize the device for specific applications. Such devices can be used for multi-spectral and hyperspectral imaging, for chemical analysis, and in surveillance and reconnaissance.

ACKNOWLEDGMENTS

The research described in this paper was supported by the SSC San Diego In-house Laboratory Independent Research Program and in part by the Center for Space Microelectronics Technology, Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (NASA).

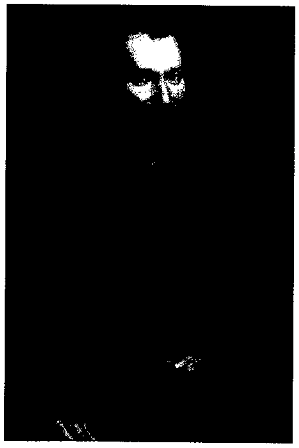
AUTHORS

Randy L. Shimabukuro
Ph.D. in Applied Physics, University of California at San Diego, 1992
Current Research: Microsensors; photonic devices; optoelectronics; microelectromechanical systems (MEMS).

Ayax D. Ramirez
MS in Physics, San Diego State University, 1991
Current Research: Microsensors; photonic devices; optoelectronics; microelectromechanical systems (MEMS); laser applications.

Michael G. Lovern
BS in Electrical Engineering, University of Arizona, 1985
Current Research: Optical target characterization; advanced optics and detectors; laser systems and applications.

Yu Wang
Ph.D. in Physics, University of Toledo, 1992
Current Research: Optoelectronic devices.



Stephen D. Russell
Ph.D. in Physics, University of Michigan, 1986
Current Research: Microsensors; photonic devices; optoelectronics; microelectromechanical systems (MEMS); laser applications.

REFERENCES

1. Levin, D. A. 1997. "Modeling of Optical Target Characterization from High Temperature Hypersonic Flows," unpublished proposal.
2. Slack, M. and A. Grillo. 1985. "High Temperature Rate Coefficient Measurements of CO+O Chemiluminescence," *Combustion and Flame*, vol. 59, p. 189.
3. Wang, Y. 1995. "Voltage-Induced Color-Selective Absorption with Surface Plasmons," *Applied Physics Letters*, vol. 67, p. 2759.
4. Wang, Y., S. D. Russell, and R. L. Shimabukuro. 1997. "Surface Plasmon Tunable Filter and Spectrometer-on-a-Chip," *Proceedings of SPIE*, vol. 3118, p. 288.
5. Wang, Y., S. D. Russell, and R. L. Shimabukuro. 1998. "Electronically Tunable Mirror with Surface Plasmons," *Proceedings of SPIE*, vol. 3292, p. 103.
6. Wang, Y. 1977. "Electronically Tunable Color Filter with Surface Plasmon Waves," *Proceedings of SPIE*, vol. 3013, p. 224.
7. Raether, H. 1980. *Excitation of Plasmons and Interband Transitions by Electrons*, monograph in series: *Springer Tracts in Modern Physics*, vol. 88, Springer-Verlag, Berlin, Germany.
8. Wang, Y. and H. J. Simon. 1993. "Electrooptic Reflection with Surface Plasmons," *Optical and Quantum Electronics*, vol. 25, p. S925.

9. Schildkraut, J. S. 1988. "Long-Range Surface Plasmon Electrooptic Modulator," *Applied Optics*, vol. 20, p. 1491.
10. Caldwell, M. E. and E. M. Yeatman. 1992. "Surface Plasmon Spatial Light Modulators Based on Liquid Crystal," *Applied Optics*, vol. 31, p. 3880.
11. Wang, Y. 1996. "Surface Plasmon High Efficiency HDTV Projector," U.S. Patent #5,570,139.
12. Russell, S. D., R. L. Shimabukuro, and Y. Wang. 1998. "Transmissive Surface Plasmon Light Valve," U.S. Patent #6,122,091.
13. Ramirez, A. D., S. D. Russell, and R. L. Shimabukuro. "Resonance Tunable Optical Filter," Patent Pending, Navy Case No. 79,095.
14. Shimabukuro, R. L. and S. D. Russell. 1993. "Microelectronic Photomultiplier Device with Integrated Circuitry," U.S. Patent #5,264,693.
15. Shimabukuro, R. L. and S. D. Russell. 1994. "Multilayer Microelectronic Photo-multiplier Device," U.S. Patent #5,306,904.
16. Fossum, E. R. 1995. "Low Power Camera-on-a-Chip Using CMOS Active Pixel Sensor Technology," *1995 Symposium on Low Power Electronics*, 9 to 10 October, San Jose, CA.

❖

Knowledge Base Formation Using Integrated Complex Information

Douglas S. Lange
SSC San Diego

ABSTRACT

An intelligence support system has been developed using open hypermedia architecture. This approach integrates information from distributed disparate sources into a knowledge base. A public interface supports access by external applications. Filtering and change detection functions have also been implemented. The approach has shown promise in multiple domains, indicating possible wide application. This paper discusses the principles of the hypermedia framework for this system and how these principles may influence command, control, communications, computers, intelligence, surveillance, and reconnaissance (C^4ISR) systems in general.

INTRODUCTION

Command and control involves three fundamental processes that fit together in a tight cycle. Situation analysis provides the context on which to act. Decisions are made based on analysis results. These decisions constitute planned movements, engagement orders, and many other possible actions. Decisions must be communicated to those who are to carry out the actions. The results of these actions are observed as part of a new situation analysis.

As command, control, communications, computers, intelligence, surveillance, and reconnaissance (C^4ISR) systems have evolved, system integration has been the general theme. Stand-alone systems, each with its own database, were first interfaced to allow some data transfer. Data management schemes provide some consistency among databases and operational units. System federation gradually allowed multiple applications to run on users' workstations, preventing the need for specialized hardware and support software for large numbers of individual systems. The current state of system integration not only allows multiple applications to share hardware, operating system, and network platforms, but also uses a layered service architecture that eliminates redundancy of some capabilities.

The evolution of system integration has broadened the stovepipes that were so narrow in previous system generations. The resulting view is of a few broad systems made up of many small applications, any of which may be accessible through the user's workstation. Some applications work on common data managed through centralized services. Many data categories still form separate stovepipes since they are maintained in separate data repositories because of their differing technical natures and programmatic backgrounds. Users must associate the tactical situation shown in one application with the results of a logistical query conducted through another application.

Information Complexity

The focus on systems integration ignores the true goal in decision support. Information is of ultimate value to the decision-makers. Integrating the information is the next step. Unlike data-warehousing applications, military information is not just collecting and crunching sales and inventory figures from various branch offices. The military environment is complex. The variety of concepts, events, and situations that can be described subjectively or measured and reported objectively is probably limitless. No ontological study can *a priori* determine all of the possible data types needed to describe the military environment. Therefore, bringing all data into a relational or object database will not completely accomplish information integration.

Pattern of Analysis

In researching the requirements for an intelligence support system for the U.S. Defense Intelligence Agency (DIA), a pattern of analysis was uncovered that was common to those used in some other domains. The primary feature of this pattern is that an analyst's role is to create associations among existing data. Analysts rarely create data, but search, filter, and review all available information. As they do, they form networks of related information [1].

DIA intelligence analysts spend some of their time building up a private model of their area of expertise. They spend the rest of their time responding to queries from DIA's various customers. The responses are typically linear essays. Analysts also periodically produce background reports on particular matters of interest. These reports also take a strictly linear, book-like form, even when delivered over a computer network.

Analysis of the current approach yielded the following problems:

- Products were static or updated using a paper publishing schedule.
- Customers with local information have no mechanism to share it with others.
- Only a particular question was answered, even if it was not the correct question.
- Analyst turnover causes a large loss of knowledge.

As a result of these insights, work was initiated to find a way of recording the knowledge built by the intelligence analyst and communicating this knowledge to intelligence consumers. The goal was to move away from the linear essay to a more collaborative communications method. This method would allow for continuous update of the knowledge jointly held between the intelligence agency and its customers.

Recording Decisions

Decisions also take the form of associations among data or information elements. A classic example may be the order for a surface combatant to engage a hostile aircraft. The decision-maker did not create the aircraft or the positional and attribute data known about that aircraft. Likewise, the decision-maker did not generate the information related to the surface combatant. The value added by the decision-maker is that an engagement relationship (perhaps with other amplifying information) should exist between the two.

As the data on the two combatants changes, the association must be reviewed, but is not necessarily invalidated. Likewise, a reversal of the decision changes the relationship among the combatants, but does not change any individual data. This fundamental distinction between the structural representation of the associations among concepts or real-world objects and the content that describes them is common between the knowledge created by analysts and decision-makers.

HYPERMEDIA ARCHITECTURES

Hypermedia systems automate the management of information that is structured as described previously. Such systems provide the capability to work with a wide variety of data, while using the powerful information available through the structures created by the connections made among the various data items [2]. Hypermedia accurately records information, but its non-linearity allows the reader to access information in ways that the author did not necessarily expect. Users of analysis results can make new discoveries from the same body of data [3]. Likewise, distribution of responsibilities in a large command and control environment is aided by ensuring that not all uses of the data must be preconceived, though accurate representation of constraints is essential.

The basic features of most hypermedia systems are as follows:

- *Node.* A node is an object that represents a document or some other media element.
- *Link.* Links create relationships among nodes.
- *Anchor.* Anchors connect nodes to the actual media that make up their content.

Open Hypermedia

From 1987 to 1991, researchers noted that the hypermedia systems did not support the needs of collaborative work groups and could not be integrated into computing environments used in large enterprises [4 and 5]. Requirements were found for hypermedia systems that were not addressed. These requirements included the following:

- Interoperability to access and link information across arbitrary platforms, applications, and data sources.
- Link and node attributes to record the author of a link, what the permissions are for the particular link or node, and other management information.
- Anchors that allow attachment to the exact data desired.
- Link types to provide more information about the meaning of a particular link and what functions the link is intended to support.
- Public and private links to support collaborative environments.
- Templates for automating routine analysis tasks.
- Navigational aids that can act as filters and supply powerful querying mechanisms.
- Configuration control so that information important to the analysis effort can be developed and managed in hypertext.

To address these requirements, open hypermedia systems evolved. Open hypermedia systems have been defined as those that exhibit the following characteristics [6]:

- A system that does not impose any markup on the data. By marking up data to create hyperlinks, the data are changed, making the data inaccessible to systems that cannot handle the markup.
- A system that can be integrated with any tool that runs under the host operating system. This can be extended to mean a system that can be integrated with distributed object environments.
- A system in which data and processes may be distributed across a network and hardware platforms.

- A system in which there is no artificial distinction between readers and authors. This requirement is quite important for systems supporting analysis.
- A system to which new functionality can be easily added.

Since analysts and decision-makers are simultaneously readers and authors of node contents and links, these characteristics are vital in an information support environment. Likewise, the ability to link objects without changing them is critical. The information linked together by the analysts may be coming from other applications and databases integrated with the hypermedia system. These applications will not understand changes imposed on the data to support linking. The links must be separated from the content. This separation is the basic premise of an open hypermedia system. It has been demonstrated in many research systems [7].

The prototypical open hypermedia system is structured as shown in Figure 1.

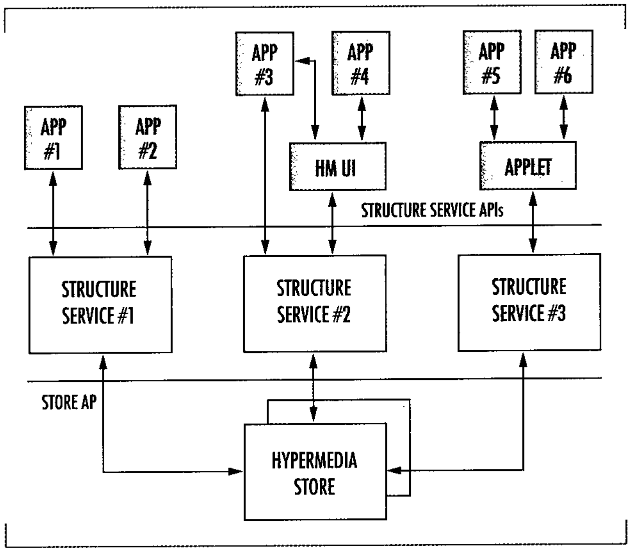


FIGURE 1. Open hypermedia architecture [9].

Graph-Based Hypermedia

Several other hypermedia system types contribute capabilities necessary to support analysis functions. Chief among these is graph-based hypermedia. Graph-based hypermedia are based on set and graph theory, providing mathematically defined filter, search, and navigation methods. This category of hypermedia also includes human–computer interaction methods featuring graphical depictions of the hypermedia.

The idea of a schema made of node and link types provides the basis for much of this method's power [8]. The relationships among schema types and between schema entries and the instances created in the hypermedia closely mirror the relationships in object-oriented design.

One result of the typing found in graph-based hypermedia systems is that the resulting hypermedia forms a semantic network. Semantic networks are used to model concepts and real-world situations, making them a natural tool for modeling a tactical situation or the results of intelligence analysis.

Another result is that sophisticated filtering mechanisms can be defined. Graph-based hypermedia provide the concept of a perspective. A perspective contains three elements. The first element is the perspective pattern. A perspective pattern is a hypergraph that is a subset of the schema hypergraph. The second element is a filter, which is a constraint on the instance set. The filter may constrain either through the node and link attributes, or the content attached through the anchors. Finally, a subset of the instance set satisfies both of the constraints.

HYPEROBJECT PROCESSING SYSTEM

The design of the HyperObject Processing System (HOPS) inherits features from both open hypermedia systems and graph-based hypermedia systems. Some modification to the established research architectures was required to support analysis of the kind performed by DIA. These same modifications would appear to be important for related C4ISR systems.

General Architecture

HOPS follows the open hypermedia form with the architecture shown in Figure 2.

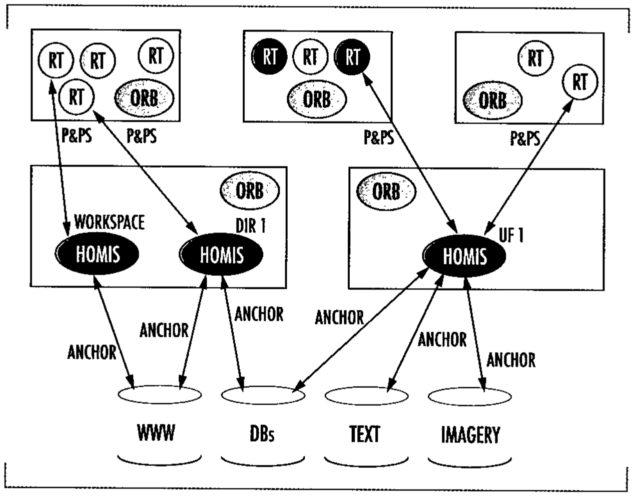


FIGURE 2. HyperObject Processing System.

In this figure, circles labeled "RT" represent runtime applications supporting a user directly or automated processing. "HOMIS" (HyperObject Multimedia Information Systems) are modified multimedia information systems [8] that can handle hypergraphs rather than simple graphs [10]. HOMIS function as structure servers, as called for in generic open hypermedia systems; however, they provide graph-based hypermedia functions. Each HOMIS has a schema and instance set. Perspectives ("P" in Figure 2) and filters can be defined, and graph-based navigation interactions are possible. "ORB" represents an object request broker, in our case, supporting Java Remote Method Invocation. Object request brokers allow the system to be distributed over multiple platforms.

Unique Hypermedia Features

Most hypermedia systems found in research literature work with information spaces constrained by either the level of diversity and quantity of the information, or by restrictions on the structure of information, or by limited change of the underlying data. Several aspects of HOPS are unique among hypermedia systems. The features are necessary to allow HOPS to handle the dynamic unbounded nature of military information integration.

Multiple Anchors

The middle layer of HOPS holds the semantic network. Classical hypermedia systems use a node to represent a piece of media and anchor to a single media element to provide content. A semantic network forms that describes the relationships among media elements rather than the tactical situation. To remedy this problem, HOPS uses multiple anchors per atomic node. Use of multiple anchors allows the nodes to define concepts or real-world objects and allows the links to represent relationships among them rather than relationships among the content elements.

Large Open-Ended Schema

Schemas imply an ability to predict all the types of information to be used and the entire range of associations that will exist among the elements. In some domains this is possible, but not in the military information domain [1]. An example can be demonstrated in terms of exercise plans. During Tandem Thrust 97, one of the primary requirements concerned protecting the Great Barrier Reef. Environmental mitigation strategies and environmental reports are not typically found in the command and control systems of our armed forces. There will always be unpredicted situations in warfare and military exercises. Information systems must adapt on the fly to allow analysts and decision-makers to see and interpret information and record and inform regarding decisions. The HOPS design allows users to include information not accounted for in the schema through the object-oriented method of deriving all nodes and links from common ancestors. This allows users to bypass rules in the schema and connect nodes and links in ways not previously predicted. The user or an administrator can then update the schema on the fly to allow autonomous tools to process the information more easily.

Analysis schemas and instance sets can become quite large. The problems modeled are quite complex. The size of the schema represents the complexity of the model while the size of the instance set represents the quantity of information. Consumers of the analysis model must filter both in terms of the complexity and in terms of the size of the knowledge base that they work with to avoid being overwhelmed. HOPS allows this capability through adaptations of the graph-based hypermedia concepts of perspective patterns and filters [10]. Perspective patterns allow the user to limit the kinds of information being worked with, while filters focus attention on information with particular content.

Link and Anchor Integrity

When important decisions are being made based on the information presented, error is less tolerable than in our daily workings with the World Wide Web. Anchored content must not disappear unexpectedly. Likewise, if content changes, the model must be re-evaluated to determine if it is still valid. The typed links of the storage layer must also be carefully managed to prevent dangling links. HOPS accomplishes these goals by caching anchored content and providing periodic checks using an autonomous change detection agent. Agents used for this purpose can use whatever rules suit the application.

Link Equality

Although hypermedia relies on associations between elements for its character, many interaction techniques found in research literature still focus on the content (e.g., string matching filters and searches, searches on images). Links are primarily used for navigation. This may be because, in many applications, links are addresses used to point to more information, or typed paths to get to related nodes. Since the primary value added by intelligence analysts and decision-makers is found in associations among elements, authors and readers of the products will want the ability to interact with typed links in ways other than simply using them for navigation. They themselves provide critical information. HOPS handles this by making links special types of nodes, allowing all the mathematics of filtering, searching, and browsing to work on links. [10].

Framework

HOPS itself is not a command and control system or an analysis system. HOPS is a hypermedia framework designed to support analysis and to provide some generic applications for interacting with the hypermedia. HOPS is intended to be used by adding domain-specific applications along with an initial schema to create an analysis system of the type needed. Such work is in support of DIA's mission.

In the Military Operations in the Built-up Areas project, HOPS was integrated with the Lightweight Extensible Information Framework (LEIF) to provide geographic and temporal views of the hypermedia.

An intelligence product creation wizard and intelligence-specific anchors were also used. Together with the generic applications within the framework, users have a variety of ways to work with the information.

PROSPECTS FOR INFORMATION INTEGRATION

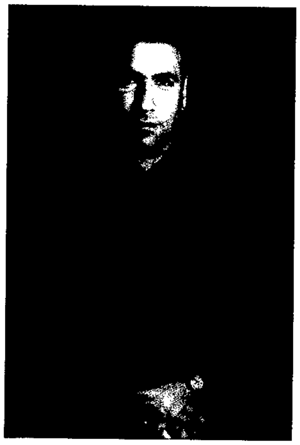
Hypermedia systems hold promise for information integration. Any number of decision-support tools can access the semantic network formed of the associations and nodes. Decision-makers can have access to all the information they need because the hypermedia can be made from information elements from all available systems. While the semantic network is serving higher level decision tools, the content is left untouched, and is still accessible by those tools that interact directly with content databases.

Beyond executing applications from a single workstation, integrated information could provide decision-makers with a competitive advantage. An integration method that brings the information into a semantic network can allow meaningful access to human beings and autonomous agents. The goal of command and control systems should be to integrate information rather than just the applications. Architecture such as that used for HOPS, centered on the structure of information, can accomplish this goal. Military plans, tactical situations, and their interaction can be described using hypermedia-induced semantic networks.

REFERENCES

1. Lange, D. 1999. "Hypermedia Potentials for Analysis Support Tools," *Proceedings of Hypertext '99*, Association for Computing Machinery (ACM), pp. 165–166.
2. Nurnberg, P., J. Leggett, and E. Schneider. 1997. "As We Should Have Thought," *Proceedings of Hypertext '97*, ACM, pp. 96–101.
3. Nielson, J. 1990. *Hypertext and Hypermedia*, Academic Press, San Diego, CA.
4. Halasz, F., 1987 "Reflections on NoteCards: Seven Issues for the Next Generation of Hypermedia Systems," *Proceedings of the ACM Conference on Hypertext.*
5. Malcom, K., S. Poltrock, and D. Schuler, 1991. "Industrial Strength Hypermedia: Requirements for a Large Engineerng Enterprise," *Proceedings of the Third ACM Conference on Hypertext.*
6. Davis, H., W. Hall, I. Heath, G. Hill and R. Wilkins. 1992. "Towards an Integrated Information Environment with Open Hypermedia Systems," *Proceedings of Hypertext 1992*, ACM, pp. 181–190.
7. Wiil, U. 1997. "Message from the OHSWG Chair," Open Hypermedia Systems Working Group Web Site, http://www.ohswg.org/intro/chair.html (December).
8. Lucarrela, D. and A. Zanzi. 1996. "A Visual Retrieval Environment for Hypermedia Information Systems," *ACM Transactions on Information Systems*, vol. 14, no. 1 (January), pp. 3–29.
9. Wiil, U. and P. Nurnberg. 1999. "Evolving Hypermedia Middleware Services: Lessons and Observations," *Proceedings of the 1999 ACM Symposium on Applied Computing*, pp. 427–436.
10. Lange, D. 1997. "Hypermedia Analysis and Navigation of Domains," Master's Thesis, Computer Science Department, Naval Postgraduate School, Monterey, CA.

❖



Douglas S. Lange

MS in Software Engineering, Naval Postgraduate School, 1997

Current Research: Software generation; knowledge bases; enterprise architectures.

A Real-Time Infrared Scene Simulator in CMOS/SOI MEMS

Jeremy D. Popp, Bruce Offord, and Richard Bates
SSC San Diego

H. Ronald Marlin and Chris Hutchens
Titan Systems Corporation

Derek Huang
Advanced Analog VLSI Design Center

ABSTRACT

A 64 x 128 real-time infrared (RTIR) complementary metal-oxide semiconductor (CMOS)/ silicon-on-insulator (SOI) scene generation integrated circuit (IC) is described. The RTIR IC offers real-time dynamic thermal scene generation. This system is a mixed-mode design, with analog scene information written and stored into a thermal pixel array. The design uses micro-electro-mechanical sensors (MEMS) in conjunction with SSC San Diego's 0.8-µm CMOS/SOI process to develop a RTIR IC scene generator.

INTRODUCTION

The objective of the real-time infrared (RTIR) project is to develop a reliable prototype infrared (IR) test set for use in calibration and testing of IR systems, including built-in-test to ensure the real-time reliability of IR sensing systems. The potential of RTIR as built-in-test equipment (BITE) is to improve the reliability of IR sensors, thus lowering the overall system cost of operation. Infrared scene simulators that use bulk complementary metal-oxide semiconductor (CMOS)/micro-electromechanical systems (MEMS) have been reported previously [1]; however, this work uses silicon-on-insulator (SOI) as the starting material. The MEMS area is scaled down to create higher density pixel arrays, with low leakage at higher temperatures.

DESIGN

The integrated circuit (IC) consists of a data input block, address write control, and pixel-specific electronics including a microheater suspended over a micromachined cavity in the silicon substrate. The display IC consists of an array of 64 x 128 thermally isolated, resistive emitters. The thermal pixel array (TPA) elements have response times less than a millisecond, making them suitable for real-time scene simulation. The pixel cell contains a resistive heater element (or infrared emitter), a storage capacitor, pixel drive transistors, and switches (Figure 1). The user digitally specifies a specific row and column and then writes a pixel voltage to the desired cell via the analog multiplexer (MUX). The infrared pixel array IC is designed for use with a computer or an electronic controller to service or update the real-time images. The computer sends gray-scale scene data to the pixel array in the form of voltages, which the TPA displays as a gray-scale image. The computer controls digital row and column address lines and writes the analog inputs via a digital-to-analog converter (DAC) to the RTIR IC. The voltage magnitude reflects the desired IR intensity of the pixel element, thereby achieving the gray-scale levels. After writing to the pixel, the desired voltage is stored dynamically by Chold, producing the desired IR pixel intensity while the remaining pixels are updated. The pixel electronics of the array are designed to exploit the low leakage properties of SOI during high-temperature operation. Voltage droop is the greatest problem affecting pixel dynamic range and accuracy. Droop is primarily a result of the leakage currents through

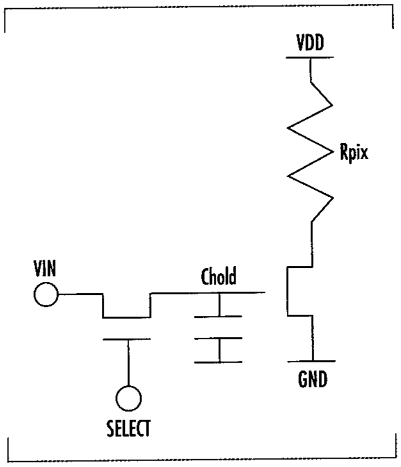


FIGURE 1. Pixel schematic: the drive transistor is a BTS device; the access transistor is an HGATE device.

the pn junctions of the sampling switch and secondarily a result of excessive channel leakage. As an option to further reduce droop, the designer can place a compensation pn junction by using half a negative-channel metal-oxide semiconductor (NMOS) transistor at the hold capacitor node.

FABRICATION

CMOS/MEMS technology is used as a technique to thermally isolate the infrared emitter microstructures from the substrate after the CMOS processing is completed. SSC San Diego's 0.8-μm CMOS partially depleted SOI process was selected to fabricate the array of electronically addressable 20 x 20 micron emitter elements (Figure 2). The process is a single poly, double metal, salicided process with a high-value resistor option of up to 1 Kohm/square. This allows modest density arrays, and, together with the high-value silicon resistor available in the 0.8-μm process, provides lower pixel current operation. The micromachined cavity is constructed by using a silicon etchant that undercuts the desired pattern in the silicon substrate, while leaving it electrically connected to create a suspended structure/microheater (Figure 3). This pattern is created by patterning and plasma etching silicon dioxide after the CMOS passivation, thereby exposing the substrate silicon of the CMOS chip. The exposed silicon is then exposed to a tetra-methyl ammonium hydroxide (TMAH) solution, an aniso-tropic silicon etchant. The TMAH etchant was chosen because, with the addition of silicic acid, it does not attack the exposed aluminum bonding pads [2].

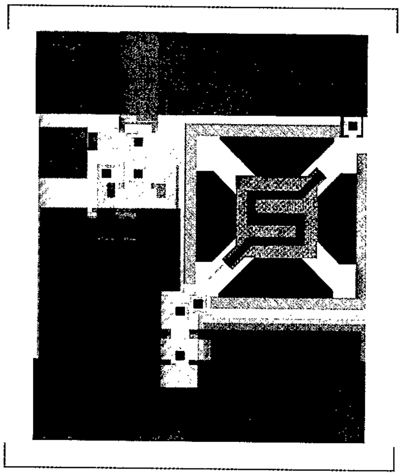


FIGURE 2. The heater element and pixel electronics layout.

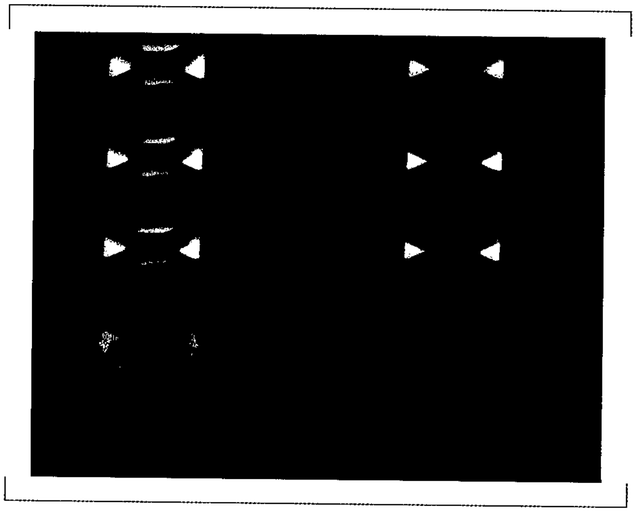


FIGURE 3. Scanning electron microscopy (SEM) of a cross-sectioned sample of suspended microheaters.

RESULTS

The thermally isolated resistor emitter has been characterized using a calibrated blackbody and adjusting for fill factor using a method described in [3]. The temperature of the emitter as a function of voltage across the resistor is plotted in Figure 4, together with the current through the resistor. A maximum temperature of 262°C is achieved at a voltage of 8.25 V and a current of 0.85 mA.

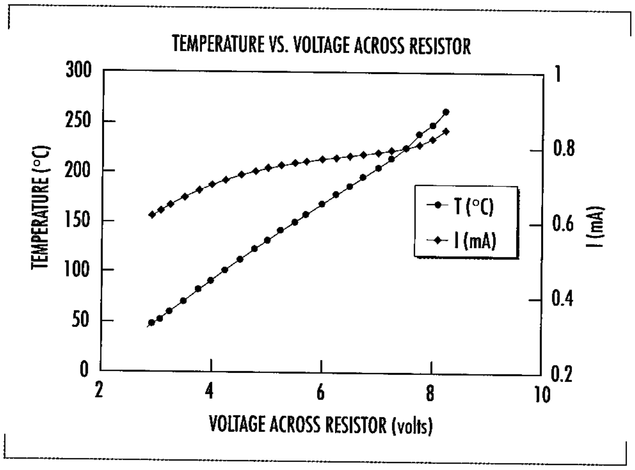


FIGURE 4. Pixel thermal response to applied voltage across resistor.

SUMMARY

A 64 x 128 scene generator RTIR IC architecture has been described with each key component discussed. A MEMS device, the TPA, is produced using CMOS/SOI technology with post CMOS process etching.

ACKNOWLEDGMENTS

This work was funded by the Office of Naval Research, Physical Science Division, Dr. Phil Abraham.

AUTHORS

Bruce Offord

BS in Engineering Physics, University of San Diego, 1985

Current Research: Very large-scale integrated (VLSI) SOI process development; novel IC design.

Richard Bates

BA in Physics, Loma Linda University, 1960

Current Research: Radiation-induced Irtran 2 absorption; polarization independent narrow channel (PINC) wavelength division multiplexing (WDM) fiber coupler fusing.

H. Ronald Marlin

BA in Physics, La Sierra University, 1959

Current Research: Infrared radiometry.

Chris Hutchens

Ph.D. in Electrical Engineering, University of Missouri, 1979

Current Research: Low-power, mixed-signal SOI CMOS and analog CMOS.

Derek Huang

MS in Electrical Engineering, Oklahoma State University, 2001

Current Research: Low-power, mixed-signal SOI CMOS.

Jeremy D. Popp

BS in Electrical Engineering, Portland State University, 1998

Current Research: Low-power, mixed-signal application-specific integrated circuit (ASIC) design; novel systems on a chip; reconfigurable computing.

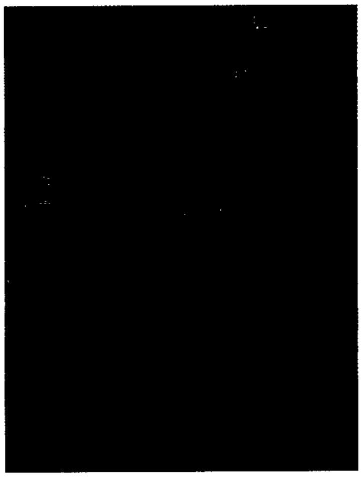
REFERENCES

1. Parameswaran, M., A. M. Robinson, D. L. Blackburn, M. Gaitan, and J. Geist. 1991. "Micromachined Thermal Radiation Emitter from a Commercial CMOS Process," *IEEE Electron Device Letters*, vol. 12, no. 2 (February), pp. 57–59.
2. Tabata, O., R. Asahi, H. Funabashi, K. Shimaoka, and S. Sugitama. 1992. "Anisotropic Etching of Silicon in TMAH Solutions," *Sensors and Actuators*, A.34, pp. 51–57.
3. Marlin, A. H. R., R. L. Bates, M. H. Sweet, R. M. Carlson, R. B. Johnson, D. H. Martin, R. Chung, J. C. Geist, M. Gaitan, C. D. Mulford, E. S. Zakar, R. J. Zeto, R. Olson, and G. C. Perkins. 1997. "Real-time Infrared Test Set: Assessment and Characterization," *SPIE*, vol. 3084.

❖

3

C4ISR Systems Integration and Interoperability ■



3 C^{4}ISR SYSTEMS INTEGRATION AND INTEROPERABILITY

C^{4}ISR Involvement with the Distributed Engineering Plant (DEP)

BeEm V. Le
SSC San Diego

ABSTRACT

The Navy's requirement for interoperability between systems and Battle Groups led to the development of the Distributed Engineering Plant (DEP). The DEP Battle Group Interoperability Test (BGIT) was a combination of several Navy laboratories in which command, control, communications, computers, intelligence, surveillance, and reconnaissance (C^{4}ISR) systems were tested within the DEP. This paper focuses on C^{4}ISR integration and interoperability testing accomplished by the DEP BGIT program. It also discusses the support that C^{4}ISR systems provide the Fleet and problems found during the DEP BGIT.

BACKGROUND

In February 1998, the Fleet was concerned about interoperability failures among combat systems recently installed in deploying fleet units. These failures led to two modern combatants being tied to the pier during their Battle Group deployment. During the final 6 months before Battle Group deployment, shipboard and Battle Group "debugging" of systems consumed valuable fleet training time. In March 1998, the Chief of Naval Operations assigned to Naval Sea Systems Command (NAVSEA) the responsibility to address combat systems interoperability problems across Battle Management Command, Control, Communications, Computers, and Intelligence (BMC^{4}I)/combat systems, and to coordinate resolution with the Fleet. In April 1998, NAVSEA formed the Task Force on Combat System Interoperability to study the interoperability crisis and provide recommendations for solutions. In May 1998, the Task Force was formally tasked to determine the feasibility and cost of using a land-based Distributed Engineering Plant (DEP) to support the design, development, test, and evaluation of interoperability of Battle Force systems. In June 1998, the Task Force on Combat System Interoperability reported that the establishment of a DEP was technically possible, but organizationally difficult because of the diverse group of organizations and elements. The Task Force also stressed that a DEP is only a tool to enable good design decisions early in the acquisition process. Following the Task Force Report, the collection of government activities represented in Table 1 formed a cooperative effort known as the Navy Alliance.

TABLE 1. Navy Alliance.

Naval Surface Warfare Center/Dahlgren Division—Dahlgren, VA
Aegis Combat Systems Center—Wallops Island, VA
Naval Warfare Analysis Station—Corona, CA
Naval Undersea Warfare Center—Newport, RI
Naval Surface Warfare Center/Port Hueneme Division (PHD)—Oxnard, CA
SSC San Diego—San Diego, CA
Naval Surface Warfare Center/PHD—Dam Neck, VA
SSC Charleston—Charleston, SC
Naval Surface Warfare Center/PHD—San Diego, CA
Aegis Training and Readiness Center—Dahlgren, VA
Naval Research Laboratory—Arlington, VA
Johns Hopkins University (JHU) Applied Physics Laboratory—Laurel, MD
Naval Air Warfare Center/Aircraft Division—Patuxent River, MD
Naval Air Warfare Center/ Weapons Division—China Lake, CA

The Navy Alliance, made up of surface, air, subsurface, and command, control, communications, computers, intelligence, surveillance, and reconnaissance (C^{4}ISR) components, crosses all Navy Systems Commands (SYSCOMS). The Navy Alliance developed a proposal for the establishment and implementation of a Navy DEP. The following sections describe the DEP concept as drafted by the Task Force, and developed and engineered by the Navy Alliance. The DEP was founded on the existence of shore-based combat system sites. These combat system sites were built to replicate the hardware, computer programs, connectivity, and environment of the ship and aircraft combat systems as much as possible. The DEP extends this concept to the Battle Group level by interconnecting these combat system sites to replicate a Battle Group. Given that the DEP is founded on shore-based combat systems, understanding the DEP begins with an understanding of a basic combat system. The combat system consists of many important elements integrated to form a system.

Space and Naval Warfare Systems Command (SPAWAR) and DEP

The plan from SSC San Diego and Space and Naval Warfare Systems Center, Charleston (SSC Charleston) was to incorporate the C^{4}ISR family of systems into the DEP. This plan complemented the Battle Group/Battle Force (BG/BF) interoperability Navy Alliance proposal, but focused on implementing the DEP C^{4}ISR component. The plan also detailed the roles of major Space and Naval Warfare Systems Command (SPAWAR) participants and provided a technical approach for integration of SPAWAR test resources with the DEP.

SPAWAR's mission was to deliver integrated interoperable C^{4}ISR systems to the operational Fleet. SPAWAR had implemented an initial capability to build, integrate, test, and support systems by establishing the Systems Integration Environment (SIE), a robust engineering infrastructure that supported this evolution. The success of the DEP was also essential to horizontal integration, not only of the SPAWAR product lines, but also between Department of the Navy (DoN) combat systems and information systems. Many combat systems and C^{4}ISR integration issues (singly and collectively) existed and needed to be identified and resolved with the DEP BG/BF integration and test process. It was SPAWAR's plan that commitment and participation in DEP by SSC San Diego and SSC Charleston would more quickly identify, quantify, and resolve fleet interoperability issues. SPAWAR's first approach was to use the SIE as a DEP extension while evaluating C^{4}ISR capability. SSC San Diego and SSC Charleston would do this by adopting a management approach that complemented the Alliance approach and by levering infrastructure and resources as much as possible. SPAWAR would phase in implementation of its C^{4}ISR site to complement the DEP process.

SPAWAR is the Navy's C^{4}ISR product and service provider, supplying advanced information systems technology to the Fleet. Programs such as the Joint Maritime Communications System (JMCOMS), Automated Digital Network System (ADNS), Global Command and Control System–Maritime (GCCS–M), Information Technology for the 21st Century (IT-21), and Navy Wide Intranet (NWI) are initiatives that are critical to the implementation of network-centric warfare. SPAWAR is initially integrating command resources to provide a virtual environment for C^{4}ISR development and testing initiatives around the globe. SPAWAR provides integrated information hardware and software systems to the Navy, other branches of the military, other agencies of the federal government, and prospective nations. The command organizational structure has three fleet-focused "Pillars"—Engineering, Installations, and Operations.

Since technology and systems change about every 16 months, training sailors and Marines on new technology becomes paramount. By focusing on deploying battle and amphibious-ready groups, SPAWAR works to ensure that new capabilities are provided to fleet units likely to need them the most—deploying Battle and Amphibious Ready Groups. SPAWAR 05 sets goals for systems engineering and for the use and management of the SIE to reduce risk, measure results, and ensure delivery of tested and validated capability to the Fleet. SPAWAR 051 is the systems engineer responsible for the development of end-to-end C^{4}ISR systems designed to provide required capabilities for each deploying Battle Group. SPAWAR 053 acts as the primary manager/test directorate for complex highly integrated C^{4}ISR integration test and evaluation. SPAWAR 053 establishes and maintains the test and evaluation processes, policies, and test infrastructure, including the SIE for the claimancy. These factors are tailored to fit specific program needs. Because the complexity of the program and its requirements vary, the management structure must have varying depth. SPAWAR 053 tailors the integration test organization to fit the complexity of each program. As a major player in the Alliance, SPAWAR 053 is a member of the Technical Advisory Board, the Systems Engineering Group (SEG), the Network Engineering Group (NEG), and the Collaborative Engineering Group (CEG). NAVSEA is assigned central responsibility to address BMC^{4}I/Combat Systems interoperability problems within the SYSCOMs/Program Evaluation Offices (PEOs) and to coordinate resolution with the Fleet.

ACCOMPLISHMENTS

The first Battle Group that SPAWAR participated in was USS *Dwight D. Eisenhower* (CVN 69) (IKE) (Figure 1). During IKE BGIT, SSC San Diego and SSC Charleston accomplished the following:

- Executed limited Y2K testing between C^{4}ISR systems and combat systems in accordance with the Navy Y2K Master Plan
- Added the ability to test a mix-match of real-time and non-real-time tracks
- Added the ability to mix live/simulated C^{4}ISR tracks
- Added the limited ability to test joint C^{4}ISR assets
- Added the ability to test C^{4}ISR interfaces to several Naval Air Systems Command (NAVAIR) platforms (E2-C, F14D (Joint Tactical Information Distribution System [JTIDS]), F18 (Multifunction Information Distribution System [MIDS]), P-3, and S-3 aircraft)
- Developed/incorporated initial Common Simulation (SIM)/Stimulation (STIM) capabilities required to test C^{4}ISR systems
- Developed/incorporated initial Data Extraction (DX)/analysis capabilities to test C^{4}ISR systems
- Led efforts to enhance and implement full collaborative engineering capabilities for the Alliance
- Provided leads in C^{4}ISR systems engineering functions in the DEP
- SPAWAR leveraged SIE test requirements and assets to address DEP goals during IKE BGIT
- Established an interface between SPAWAR C^{4}ISR SIE and DEP, which replicated the ship configurations for the Automated Digital Network System (ADNS), GCCS–M, and the Officer in Tactical Command Information Exchange Subsystem (OTCIXS) for the IKE BGIT

- Planned and conducted a Battle Group Interoperability (BGI) Test Program that included C4ISR, combat systems, and several select "multi-source inputs"
- Supported the Navy Y2K Master Test Plan Level 2 and Level 3 test for C4ISR systems that interfaced to combat systems
- Supported the development of a "common" SIM/STIM C4ISR component for use in the DEP and SIE SIM/STIM environment.

Besides the IKE BGIT, SPAWAR has been a participant in the USS *George Washington* (CVN 73), USS *Abraham Lincoln* (CVN 72)/USS *Harry S. Truman* (CVN 75), USS *Constellation* (CV 64)/USS *Enterprise* (CVN 65), and USS *Carl Vinson* (CVN 70) BGITs. During these BGITs, several technical reports were written to document fleet findings for the C4ISR systems, particularly GCCS–M and Common Operational Picture (COP) Sync Tools (CST). These problems have been documented and reported to the Fleet and Program Office for correction.

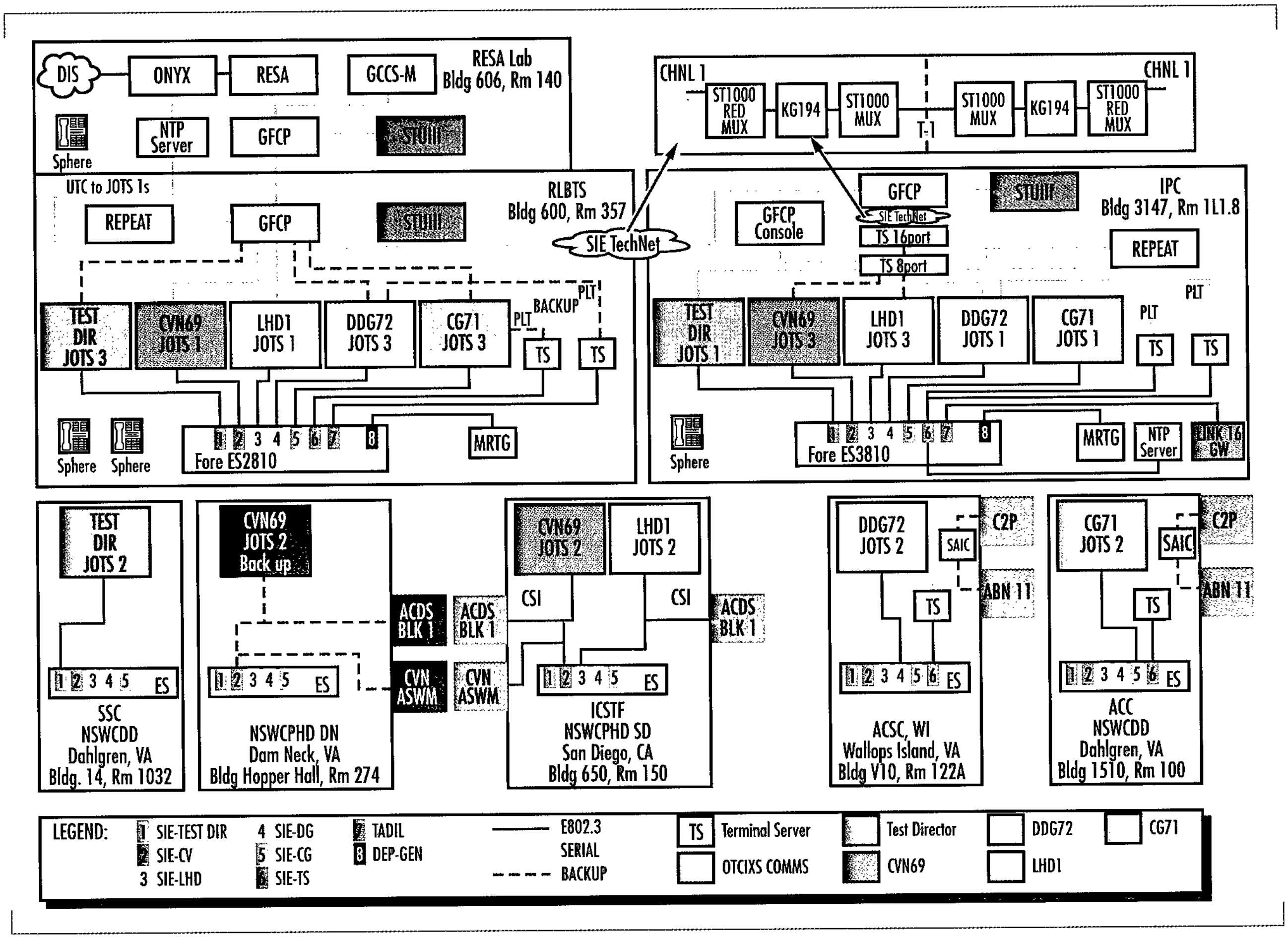


FIGURE 1. IKE DEP/SIE architecture.

LOOKING FORWARD—THE FUTURE

For the intermediate future, SPAWAR is planning to participate in USS *John F. Kennedy* (CV 67) BGIT, which is scheduled in June and July

2001. For this BGIT, GCCS–M will interface with the Advanced Combat Direction System (ACDS) Block 1 (two-way Combat System Integration [CSI] interface) and will interface with the Air Defense Systems Integrator (ADSI).

Looking ahead to FY 2002 and beyond, SPAWAR is planning to support and can include Joint Systems and Coalition Systems into the DEP. The overall focus of the original DEP Systems Engineering effort was to set up a disciplined and robust systems engineering process that leads to the development of a more interoperable joint force and the development of the DEP required to support that process. SPAWAR's system engineering process supports the concept in which the BF is the warfighting system rather than an individual platform. SIE offers a proven capability to build and test valid C^{4}ISR architectures, which represent the complex operational C^{4}ISR environment. The C^{4}ISR SIE will further develop the DEP's ability to support overall force requirements to have interoperability "engineered-in." The direct interfaces between C^{4}ISR and combat systems are limited today; however, highly integrated C^{4}ISR systems on the other side of the direct interface system (e.g., GCCS–M) provide multi-source inputs that are fused together, providing vital information to the warfighter. Interoperability testing requires that many components besides the direct interface system be tied into the test architecture. Network-centric warfare and NWI will provide important timely information, extending the battlespace and supporting advanced mission planning. SPAWAR's commitment to the DEP will also support future efforts, including a closer integration of real-time and non-real-time command and control (C^2), development of a common information base for C^2, and integration of the Tactical Digital Information Links (TADILs) into the common backbone. A valid C^{4}ISR architecture has elements that operate at UNCLASSIFIED, SECRET, and Sensitive Compartmented Information (SCI) classification levels. All three are crucial for accurate integration and valid interoperability testing for BG C^{4}ISR architecture and the integrated network security.

The original DEP effort was designed to support the important interoperability requirements of:

- A common tactical picture across all force elements
- The control and coordination of engagements at the force level
- Force-level planning

SPAWAR's specific goals, with other SYSCOMS, are to add the following important interoperability requirements of C^{4}ISR:

- A common operational or tactical picture across all force elements
- Inclusion of the intelligence, information warfare (IW), cryptologic, and mission planning elements of BMC^{4}I
- Inclusion of the meteorological, navigation, logistics elements of BMC^{4}I
- Ability to simulate the NWI and Global Networked Information Enterprise (GNIE)
- Inclusion of real and simulated C^{4}ISR networks (e.g., radio frequency [RF] and Internet Protocol [IP] networks)
- Integration of real-time and non-real time C^2 to include an integrated information base (IIB)

- Integration of the TADIL data into the common backbone
- GCCS–M for Submarine Combat Systems
- The COP Test will verify the capability to provide a common operational picture environment for interoperability testing. Several protocol scripts will be used to drive multiple SIMs/STIMs at various DEP sites. Data will be recorded. Track databases from C^{4}ISR C^2 system base lines will be compared to ensure replication of known and/or expected performance.
- Link capabilities related to C^{4}ISR will be tested to ensure the C^{4}ISR DEP's capability to test interoperability. These capabilities include ADSI, multi-TADIL capability (MTC), and GCCS–M Tactical/Mobile, Coast Guard Link 11, and other related capabilities.

❖

BeEm V. Le

BS in Electrical Engineering, Bradley University, 1987; BS in Mathematics, Bradley University, 1987

Current Research: Interoperable C^{4}ISR systems; IT-21.

The Over-the-Horizon Targeting (OTH-T) Program and the Reconfigurable Land-Based Test Site (RLBTS) Laboratory

Gary E. McCown
SSC San Diego

ABSTRACT

This paper focuses on command, control, communications, computers, intelligence, surveillance, and reconnaissance (C^{4}ISR) integration and interoperability testing accomplished by the Over-the-Horizon Targeting (OTH-T) program and the support that the OTH-T program provides the Fleet, including technical expertise afloat and ashore for submarines, surface, and land-based components. Test scalability from recent small-scale tests such as Web replication (Fleet-requested) to large-scale projects such as the Distributed Engineering Plant (DEP) are also discussed. This paper also addresses the Fleet Systems Engineering Team (FSET). FSET support provides system engineering to command centers and numbered fleet commanders, daily network monitoring and troubleshooting of the Officer in Tactical Command Information Exchange Subsystem/ Tactical Data Information Exchange System to Pacific Fleet/ Atlantic Fleet command centers, and data collection and analysis tools.

INTRODUCTION

The Over-the-Horizon Targeting (OTH-T) program conducts interoperability certification testing in accordance with Office of the Chief of Naval Operations instruction (OPNAVINST) 9410.5. OPNAVINST 9410.5 requires interoperability certification for new/upgraded systems to proceed to Operational Evaluation (OPEVAL). This instruction provides for configuration management, process and plan development, and requirements development for U.S. Navy and Joint interoperability testing.

To fulfill the charter of OPNAVINST 9410.5, the OTH-T program provides a virtual, global systems integration and test facility for Information Technology for the 21st Century (IT-21) command, control, communications, computers, intelligence, surveillance, and reconnaissance (C^{4}ISR) technology. This technology collects, transmits, correlates, and displays track data into a Common Operational Picture (COP) to support warfighting requirements. The common view of the battle space that the COP provides the warfighter has been applied across the spectrum of warfare missions areas; however, the technology and doctrine have changed radically in recent years and continue to change rapidly. Thus, the primary goal of the OTH-T program is to transition architectures and systems from older military standard (MIL-STD) technologies to commercial/government off-the-shelf (COTS/GOTS) technologies. Another goal of the OTH-T program is to support the integration of all C^{4}I systems into warfighting capabilities; this support included Year 2000 (Y2K) interoperability and integration testing and direct fleet support. Fleet support also includes providing technical expertise afloat and ashore through highly trained experienced Fleet Systems Engineers (FSEs) who ensure smooth integration of new C^{4}ISR capabilities during major fleet exercises and demonstrations that validate and evaluate developed portions of configurations. The OTH-T program performs integration and interoperability testing to support warfighting capabilities for MIL-STD and IT-21 COTS/GOTS equipment for submarines, surface, and land-based components. The Fleet System Engineering Team (FSET) provides system engineers to support command centers and numbered fleet commanders; Officer in Tactical Command Information Exchange Subsystem/ Tactical Data Information Exchange System (OTCIXS/TADIXS) network monitoring and troubleshooting support to Pacific Fleet/Atlantic Fleet (PACFLT/LANTFLT) command centers; data collection and analysis tools for FSEs (ashore and afloat); test coordination/direction for system integration testing; and coordination with other certification agencies.

BACKGROUND

Experiments in the 1970s showed the difficulty and problems associated with maintaining command and control across platforms with many individual platforms developing their own tactical picture and sharing that picture. The Office of the Chief of Naval Operations (OPNAV) established the OTH-T program in 1985 to address these problems. The OTH-T program was originally tasked to develop communications specifications and Battle Group Data Base Management (BGDBM). The objective of the OTH-T program is to produce a complete, accurate, timely, precise, tactical picture suitable for getting ordnance on target where all participants have access to the correct information. The OTH-T program established the Reconfigurable Land-Based Test Site (RLBTS) in 1989 to allow interoperability and integration testing. The OTH-T program is funded through OPNAV N6 and receives Operational Maintenance, Navy (OM&N) funding to support the RLBTS Laboratory and facilities. Other major sponsors include Space and Naval Warfare Systems Command (SPAWAR) PMW 157 and 165, and PACFLT.

THE RECONFIGURABLE LAND-BASED TEST SITE (RLBTS) LABORATORY

In the early 1990s, the Navy designated the RLBTS Laboratory as the lead OTH-T laboratory. RLBTS was established as a facility to support the development of tactics and procedures for targeting systems and weapons, concept demonstrations of prototype systems, and the definition of architectures intended to ease future acquisition decisions. RLBTS provides the Navy with a facility that maintains command, control, and communication systems expertise to ensure technical and scientific excellence that provides the corporate knowledge, technical networking innovation, and real-world understanding to support operationally effective fleet warfare mission area systems. The OTH-T program operates RLBTS as a full-service facility for conducting Joint Distributed Tests and Evaluations (DEP and Joint DEP) and OTH-T system integration interoperability tests and certifications. RLBTS is expandable to support command and control configurations from the platform level to the afloat/ashore Command Center level. RLBTS provides a test control center hub, network operations center (NOC), a focal point for all test data collection and analysis, a classified test environment, architecture development and validation, and network engineering to support Fleet Command Centers.

FIGURE 1. The RLBTS Laboratory.

FIGURE 2. The Tactical Analysis Section of the RLBTS Laboratory.

FIGURE 3. View of REPEAT machines and JOTS1 tactical workstations.

Figure 1 shows a combined view of Joint Operation Test Site (JOTS) workstations and the multimedia center. The large screen display (Figure 1) can be connected to any workstation and various videoteleconferencing (VTC) units. A whiteboard and VTC unit are permanently connected to SPAWAR Headquarters (SPAWAR HQ) and SSC Charleston for collaborative real-time test planning and test execution. Figure 2 shows the Tactical Analysis Section of the laboratory. These machines house the Repeatable Performance Evaluation and Analysis Tool (REPEAT) used for tactical data recording, analysis, and playback. Figure 3 shows

REPEAT machines and JOTS tactical workstations in the Tactical Data Section.

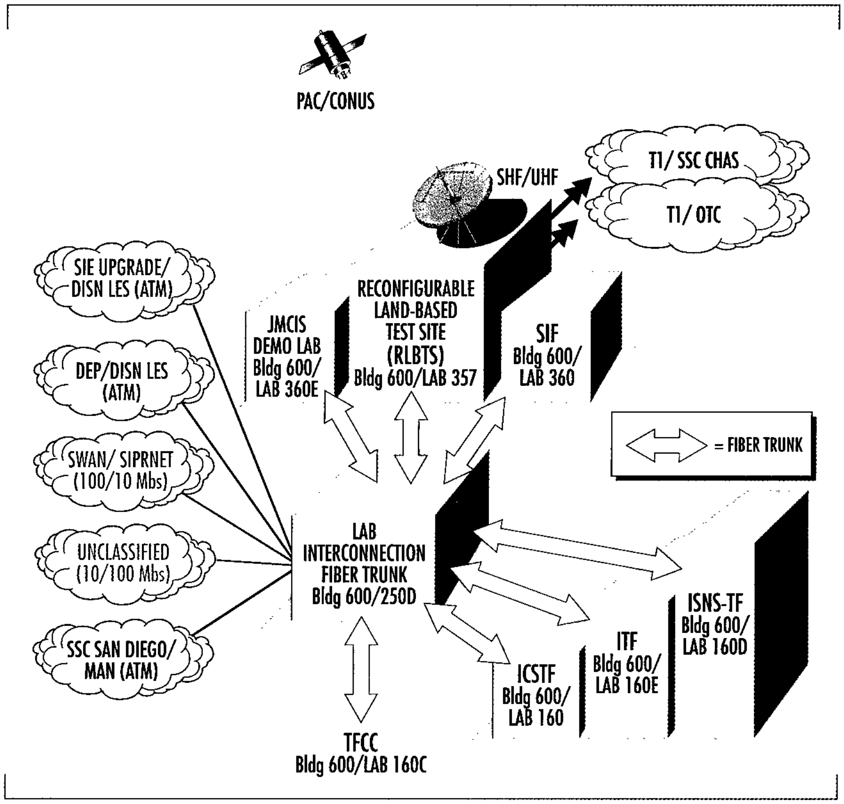

PAC/CONUS
SHF/UHF
T1/ SSC CHAS
T1/ OTC
SIE UPGRADE/ DISN LES (ATM)
DEP/DISN LES (ATM)
SWAN/ SIPRNET (100/10 Mbs)
UNCLASSIFIED (10/100 Mbs)
SSC SAN DIEGO/ MAN (ATM)
JMCIS DEMO LAB Bldg 600/ LAB 360E
RECONFIGURABLE LAND-BASED TEST SITE (RLBTS) Bldg 600/LAB 357
SIF Bldg 600/ LAB 360
= FIBER TRUNK
LAB INTERCONNECTION FIBER TRUNK Bldg 600/250D
ISNS-TF Bldg 600/ LAB 160D
ITF Bldg 600/ LAB 160E
ICSTF Bldg 600/ LAB 160
TFCC Bldg 600/LAB 160C

FIGURE 4. RLBTS networking communication and laboratory interconnection.

Network infrastructure supported by the RLBTS Laboratory (Figure 4) includes secure fiber connections to other laboratories, including the Integrated Shipboard Network System–Test Facility (ISNS–TF) (NOC); Integrated Test Facility (ITF); Integrated Combat System Test Facility (ICSTF, NAVSEA); Research, Evaluation, and Systems Analysis (RESA); Global Command and Control System–Maritime (GCCS–M), and Systems Integration Facility (SIF), with T1 and Asynchronous Transfer Mode (ATM) connectivity to SPAWAR-HQ, SSC Chesapeake, and SSC Charleston. Network connectivity also includes the Systems Integration Environment (SIE) Upgrade (T1)/Defense Information Systems Network–Leading Edge Services (DISN–LES, ATM); DEP/DISN–LES; Defense Research and Engineering Network (DREN); Ship Wide-Area Network/Secret Internet Protocol Router Network (SWAN/SIPRNET); and SSC San Diego networks. The RLBTS Laboratory also maintains a satellite communications capability to PACFLT and LANTFLT assets. RLBTS Laboratory assets include routers, packet shapers, ATM switches, firewalls, cryptos, multiplexers, and satellite simulators. Additionally, a capability to emulate the Integrated Shipboard Network System (ISNS) shipboard network has been developed by the OTH-T Program.

TESTING AND OTHER ACCOMPLISHMENTS

The OTH-T program has certified the interoperability of systems and software including OASIS, GCCS–M, Combat Control System (CCS) MK II, and the Advanced Tomahawk Weapon Control System (ATWCS). These certifications are performed annually or as new versions or software patches are developed for the Fleet to meet OPNAVINST 9410.5 requirements.

During FY 1999, the OTH-T program conducted systems integration, interoperability, and Y2K testing using the facilities of the Land-Based Test Network (LBTN), and expanded RLBTS to validate IT-21 technologies prior to shipboard installation. The OTH-T program conducted 27 tests, recommended certification of 3 systems during 59 test weeks, produced 229 documents, and provided 43 Software Trouble Reports (STRs) to program managers and system developers. OTH-T team members also participated in the DEP Battle Group Interoperability Test (BGIT) for USS *Dwight D. Eisenhower* (CVN 69) and USS *George Washington* (CVN 73), and developed and tested the COP Synchronization Tools (CST) functional requirements.

During FY 2000, the OTH-T program conducted integration and interoperability testing using the LBTN, SIE, and the IT-21 infrastructure in the RLBTS Laboratory connected to various facilities around the U.S. The OTH-T program conducted 29 tests, recommended certification of GCCS–M 3.1.2.1 and CCS MK II during 149 test weeks, produced 191 documents, and provided 91 STRs to program managers and system developers. Forty briefs were given in the RLBTS Laboratory.

The OTH-T team supported the DEP BGIT of *Eisenhower, George Washington*, USS *Carl Vinson* (CVN 70), and USS *Constellation* (CV 64) BG C4ISR configurations. The OTH-T team's participation in the test readiness reviews, test execution BGIT analysis review panels, and scheduling meetings for SPAWAR led to the identification of 24 Test Observation Reports (TORs). TORs are used to isolate problems and provide a fix or work-around recommendation.

Interoperability and Integration Testing

Specific interoperability and integration testing was accomplished for the CST segment in GCCS–M and GCCS–M version 3.1.2.1. Fifty-six STRs were recorded with 30 high, 18 medium, and 8 low. These STRs were passed to the developer and sponsor and recorded in the OTH-T database. Certification was recommended with Interim Authority to Operate (IATO) in the Fleet. As a follow-on to the certification testing, OTH-T test engineers participated in developmental testing / operational testing (DT/OT) and OPEVAL with USS *Enterprise* (CVN 65) at sea. The DT/OT demonstrated the capabilities of the GCCS–M/CST software. The OTH-T program also provided test procedures and lessons-learned reports. As this software is installed in the Fleet, the OTH-T program provides technical support and additional testing as requested by users.

Interoperability certification tests were conducted for the submarine weapons CCS MK II. Interoperability certification was recommended for the CCS MK II system. During DT, eight STRs were identified. These STRs were identified before certification, and fixes or work-arounds were implemented.

Additional interoperability/integration testing included joint testing with the Joint Interoperability Test Command (JITC). The RBLTS Laboratory participated as a node on a wide-area network (WAN) on SIPRNET testing of GCCS–M and GCCS–J (Joint). Additional participants were the Naval Center for Tactical Systems Interoperability (NCTSI) and the Defense Information Systems Agency (DISA).

Repeatable Performance Evaluation Analysis Tool (REPEAT)

The OTH-T program initiated the development of REPEAT and supports its maintenance, use, distribution, and continued development. REPEAT monitors and tests all C^{4}I synchronous and asynchronous serial devices. REPEAT monitoring and testing allows the user to conduct statistical analyses on volume and type of data, system throughput and timeliness, tactical data network loading, correlation accuracy, system data loss, common tactical picture, and comparison of data transmitted and received at various locations. Message formats that are currently supported include OTH-Gold, TACREP, TADIL-A, TADIL-B, TADIL-J, RAINFORM, Tactical Information Broadcast System/Tactical related application Data Distribution System (TIBS/TDDS), Tactical Electronic Intelligence (TACELINT), LOCATOR, Tactical Receive Equipment (TRE), Tactical Fire Direction System (TACFIRE), and Sensor Tactical Contact Report (SENSOREP). REPEAT tests OTCIXS/TADIXS/SIU/V6 interfaces. The OTH-T team provides software support, training, and upgrades. REPEAT is available in MS-DOS and Windows versions. REPEAT is currently installed and used for data analysis and recording at over 300 military sites at more than 55 commands and allied militaries. More than 300 help calls are handled each year. REPEAT software is available to all U.S. military at http://repeat.spawar.navy.mil. REPEAT provides scenario development and data/platform injection for Joint Warrior Interoperability Demonstration (JWID) exercises. During FY 2000, REPEAT supported the Global Positioning System (GPS) Inter-Service Agreement (ISA) Demonstration (sponsored by Fleet Battle Laboratory), specifically injection of GPS messages into GCCS–M. REPEAT is currently installed on many Navy platforms and is used by the Fleet to identify problems.

Test Process Web-Enabled

The OTH-T program maintains a password-protected Web site at http://otht.spawar.navy.mil to support the OTH-T team and promote process documentation, process improvement, and configuration management. The Web site allows documentation development, a tester log, engineering notes, test planning, and process documentation.

FLEET SYSTEMS ENGINEERING TEAM (FSET)

The OTH-T program supports the Fleet Systems Engineering Team (FSET), which is the main technical advisor to carrier Battle Group (CVBG)/amphibious ready group (ARG) staffs in matters related to the IT-21 architecture, associated C^{4}ISR/information operations (IO) systems, and supporting networks and infrastructures. Besides serving as a technical liaison on system management issues, the FSET also interfaces with those baseband systems that provide connectivity between the shore and shipboard networks. This connectivity includes Challenge Athena, super high frequency (SHF), Automated Digital Network System (ADNS), and other line-of-sight systems. Integrated with LANTFLT and PACFLT Commander-in-Chief (CINC) N6 organizations, the FSET also monitors all CVBG/ARG C^{4}ISR installations and liaisons with ship C^4 installation supervisors to verify that all required connectivity is in place to support tactical operations.

Coordinated with the RLBTS Laboratory, the FSE team provides system engineering support for experiments and tests that support the introduction of new SPAWAR Common Operational Picture (COP) software/hardware or system capabilities. Systems engineering will support pretest coordination, test design, installation test and coordination, and onsite support when required at remote facilities, data collection and analysis, injection of synthetic data, and post-test lessons-learned reports. FSETs provide daily support to the numbered commanders and CINC staff and their command Assist CINC in developing C^{4}I architectures and requirements. Support includes system-level support for C^{4}I non-real-time systems during BG work, Battle Group Systems Integration Test (BGSIT), and exercises (for example, Joint Task Force Exercise [JTFEX], Cobra Gold, Kernel Blitz, Tandem Thrust, and Magellan).

As representatives of SPAWAR and the Fleet CINC, the FSET ensures that deploying forces have ready access to technical experts familiar with the IT-21 architecture from an installation and operational point of view. FSET support is available upon request to major staffs throughout their deployment workup cycle. The FSET provides the ship, staff, and Battle Force an "on-scene" representative, uniquely experienced in the afloat architectures. Information on how to request FSET support is available by contacting either the LANTFLT or PACFLT program managers. In coordination with the OTH-T program, the FSET provides rapid response problem solving for issues encountered in the Fleet.

LOOKING FORWARD—THE FUTURE

With the infrastructure that has been established in the RBLTS Laboratory and connections to many other facilities from the RLBTS Laboratory, the future looks busy and full of new opportunities. These opportunities are described in the following subsections.

Multiple Large-Deck BG Interoperability Testing

Multiple large-deck BG interoperability testing ensures that multiple large-deck BGs can communicate and share data for collaborative planning, COP, and dissemination of (air) tasking orders on virtual local-area networks (VLANs), LANs, or WANs. This testing is required as a result of previous fleet observations of conflicts that involved multiple BGs converging in an operational theater with interoperability problems that forced technical experts to quickly respond to the BGs and implement work-arounds to ease operations—a costly occurence. Multiple large-deck interoperability testing of C^{4}ISR systems has never been executed in preparation for multiple large-deck contingencies.

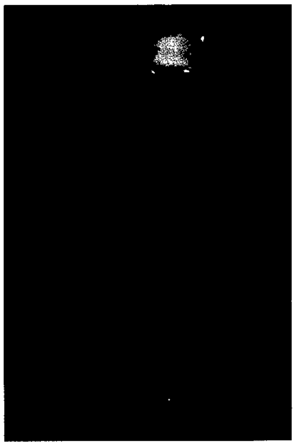
Prioritized Products List (PPL) Testing

Prioritized Products List (PPL) testing for intensified rapid-response interoperability testing is required because of changing shipboard infrastructure and networks, the use of multiple vendors, increased complexity of systems and software, an increased number of nodes/participants, increased geographic extent, and the complexity of required networks. Equipment upgrades and evolutions require more regression testing, verification, and validation to ensure and certify that new and legacy software function according to specification and interoperate with other equipment and platforms. Critical issues include WAN/bandwidth management and the extent of impact of applications when WAN bandwidth is constrained or dirty satellite conditions exist. Interoperability problems will become evident when hardware and software are installed in the Fleet and during fleet operations rather than in the laboratory. Mission capabilities will be reduced. Costs to repair problems and develop *ad hoc* fixes found in the Fleet will greatly exceed costs to identify and fix problems found in an ashore test environment. The OTH-T program with the RLBTS Laboratory is ideal for this PPL testing because of existing NOC, Integrated Shipboard Network System (ISNS), satellite simulation, and WAN facilities and expertise.

CONCLUSION

The OTH-T program will continue to provide the Fleet with high-quality products in the conduct of integration and interoperability testing of OTH-T and combat systems with tactical data exchanged over CST networks and other networks. Integration testing will include testing of GCCS–M and Combat Decision Systems (CDS) two-way interfaces. The OTH-T program will continue to support FSET integration tests and fleet test requests, horizontal integration, relevance verification, modification recommendations, and OTH-T specification maintenance to support distribution of C^4ISR systems to the Fleet, and participate in DT, OT, and OPEVAL as required. OTH-T will also provide certification testing as required by OPNAVINST 9410.5.

❖



Gary E. McCown

Ph.D. in Atomic/Surface Physics, Oregon State University, 1990

Current Work: Program Manager for the Over-The-Horizon Targeting (OTH-T) Program.

Automation in Software Testing for Military Information Systems

Jack Chandler
SSC San Diego

ABSTRACT

Software testing must be definable, measurable, consistent, and objective to be a repeatable process. This paper examines the components of the testing process, including software, hardware, the human element, and the data-collection process. It also includes a case study in test automation derived from the Defense Information Infrastructure Common Operating Environment (DII COE). A reduction in the number of human-controlled steps in the software process significantly improved test results during this case study. Automation was successful because the many different components of software were tested for compliance to a well-defined standard. Automation was straightforward because the test methodology did not require any specific assumptions about the software tested.

INTRODUCTION

This paper shows how automation can improve test results. At the beginning of this effort, a search was conducted to survey the status of automated testing. The survey revealed white papers and some commercial products that help automate testing (see Dustin [1] and Pettichord [2]). Many of the commercial products are key and cursor recorders that capture the keystrokes and cursor movements produced by test engineers during the testing process. This testing works well for testing revisions of the same product. It is not as appropriate for testing multiple pieces of software for compliance to a standard.

Dustin's paper on introducing automation to a test team states that the first phase of designing testing automation is analyzing the testing process [1]. To be of value, software testing must be a repeatable process that is definable, measurable, consistent, and objective. If the process is deficient in any of these areas, the testing will not be repeatable. This paper examines various factors in the testing process (including the human factor), describes the results of a case study on military information systems, reviews the steps required for successful automation, and provides a conclusion.

COMPONENTS OF THE SOFTWARE TESTING PROCESS

Software under Test

The most essential and basic component of the testing process is the software under test. This component cannot be changed to any great extent. The two basic categories of software under test, depending on the type of testing, are as follows: (1) testing a software product to determine whether the product is ready for release or to validate error corrections, and (2) testing multiple components of software for compliance to a standard. Both types of testing are valid; they have different requirements and different automation strategies.

Hardware for Software Testing

The second component of the testing process is the hardware platform on which the software is loaded. Improvements may be possible in this area. For example, a different hardware platform may increase the speed of software testing. Increasing the number of "seats" in use simultaneously during the testing scenario or increasing the performance of the individual "seats" is another way hardware can improve testing. The disadvantage of substituting different hardware during the testing process is the risk of testing on a non-representative platform, which would make test results questionable. To overcome this potential problem, some testing must be done on a typical user platform.

Human Testers

The third component of the testing process is the engineer or group of engineers testing the software. Test engineers make a substantial contribution to the testing process, but the possibility of human error makes them the weakest factor in the testing process. Even more important is the fact that the process is not consistent because test engineers do not consistently make the same mistake. The most dedicated and competent engineer can err under some circumstances. Thus, eliminating the "human in the loop" can significantly improve the testing process. Reducing the number of human-controlled steps can dramatically improve software testing. An example of how to reduce the human interface areas in a testing process is described below.

Data Collection

Another major component of the software testing process is the data-collection function, which often can be improved. The data-collection function often consists of the test engineer manually filling out a paper data-collection form. The test engineer will have a test notebook or a data form in which the test data are recorded. Often, the test data are re-entered into a spreadsheet or a word processor for report generation or into an e-mail message for distribution of the test results. Forms, which provide ample opportunity for errors, could be significantly improved. For example, if the form is structured in a multiple-alternative, forced-choice paradigm rather than a less-structured essay format, subjectivity can be reduced.

Another common test procedure consists of the test engineer manually filling out an electronic data-collection form. The electronic form is better than the manual form because data are manipulated only once, thus reducing transcription errors if the data are input into a report generator or an e-mail system. Using an electronic form can require investing in more hardware to support the collection. More time may be needed to fill out the forms initially, but this method saves time by reducing or eliminating the need to transpose data.

As with paper forms, the design of the electronic form is critical. One way to improve the design is to minimize the number of probable answers while still allowing all possible answers. This is done by prompting the user to consider certain likely choices while grouping other possible answers under "other" with a space to insert a comment. A periodic review of the use of the "other" category is recommended, with the objective of providing common "other" answers with specific choices of their own.

Another useful mechanism is to collect data automatically and manually by enabling software to perform the test. Efficient results are achieved by automating to the fullest practical extent the test data acquisition process. The parts of the process that do not lend themselves to automation still can be performed manually. A useful, proven procedure is to provide in the testing software a mechanism to input the manually derived test information. The "form" that is provided for collecting this manual information should be designed using the criteria discussed previously.

The most advanced and desirable phase of automation consists of collecting the testing data with a computer program that involves little or no human involvement. Only when the testing process is completely automated is a repeatable process achieved. Whenever a person manually performs a test, there is a chance that the test cannot be consistently repeated. Human beings are predictable in a group, but unpredictable individually.

Other Components

The other two major components of software testing are the education of the test engineers and the testing process itself.

TEST AUTOMATION: A CASE STUDY

Background

This section describes an example of end-to-end testing where the testing itself has been mostly automated and the areas that cannot be automated have been analyzed to reduce or eliminate subjectivity. The Department of Defense (DoD) has created the Defense Information Infrastructure Common Operating Environment (DII COE). Many DoD systems are being built using this "plug and play" infrastructure. The components of software for this system are called segments. A compliance specification has been created to enhance the "plug and play" capability of this infrastructure. This specification consists of over 300 requirements. A segment must pass at least the first 200+ requirements to be considered for inclusion in the DII COE.

DII COE compliance involves a time-consuming and human-intensive testing process. In one instance, about 18 person-hours were required to test a single, simple segment. Significantly greater test durations have been common in other cases. Compliance testing was a likely candidate for automation because it is common to all segments. The procedure that was used to automate this testing process is described below.

Document the Testing Process

In this step, the engineer will discover the current method of testing, including the acceptable test methods and the methods that are unsatisfactory. In many cases, the test engineer will be able to learn the test processes and procedure and to expose many inconsistencies in this step. The information that needs to be recorded during the test procedure is documented. From this experience, the test team learned that there were many "homegrown" solutions to the automation, a situation that had advantages and disadvantages. On the positive side, some work already had been completed. Unfortunately, these solutions were not consistent. After gleaning the currently automated processes, the test team captured the steps involved in those areas that were not automated.

The test team learned that not all DII COE requirements were tested. This was not because the untested requirements did not provide any added value, but rather to reduce the time required for testing. This prompted the test team to create a "best practices" spreadsheet in which to capture the test algorithms, not the test "programs." To provide better DII COE compliance, the test team also created algorithms for the requirements that were not being tested.

Identify Common Processes

The test team identified the common processes from the algorithm spreadsheet. These will be used later to ensure that a single testing method will be used. Too often, common testing processes are coded multiple times because engineers are unaware that some of these processes are in common use. This causes inconsistencies in the use, application, and maintenance of these processes, especially if the processes need debugging or upgrade.

Design the Automation (Software)

The first steps of the automation design consisted of gathering and defining requirements. Then, the software had to be designed to meet those requirements. The design called for an object language with a compliance engine and an individual object for each requirement. Since the DII COE software is supported on multiple platforms, the test team elected to use Java as the programming language. In theory, this was expected to decrease the inconsistencies by providing a common baseline. The test team wanted to keep the design generic to accommodate any required testing process where the requirements were structured in hierarchical levels. To pass at level 5, for example, would require that all tests in levels 1 through 4 be passed as well as all tests in level 5. With this in mind, the test team designed a compliance engine with a test manager, an applicability filter, and some common data-collection agents. We defined an interface to the compliance engine that the test objects will use. The design included a report generator and a graphical user interface (GUI) that allows the test engineer to view the data-collection form and to access the various options. These components are described below in more detail.

Because many of the DII COE requirements apply only to certain types of segments, the test team needed an applicability filter to determine the applicable tests based on the segment type. Each test object specifies to the applicability filter the segment types to which it applies. The default was specified as applicable to all segment types.

The test manager launches the test objects at the appropriate times. For example, certain tests must be run while the segment is installed, whereas others cannot run until the segment is deinstalled. The test manager also runs some data-collection agents, which also must be run at certain times. The relative timing of each test is important to specify clearly when documenting the testing process.

The data-collection agents determine information about the segment under test and that segment's effect on the underlying system. The test-unique data-collection agents (if any) are common processes that collect other data from the underlying system. This feature was included in the design, not for the specific purposes of the test team, but only for the generic case.

The questionnaire object(s) obtain additional information from the test engineer. Questionnaires are presented twice in the testing process. (Figure 1 shows the testing process.) The first time the questionnaires are presented

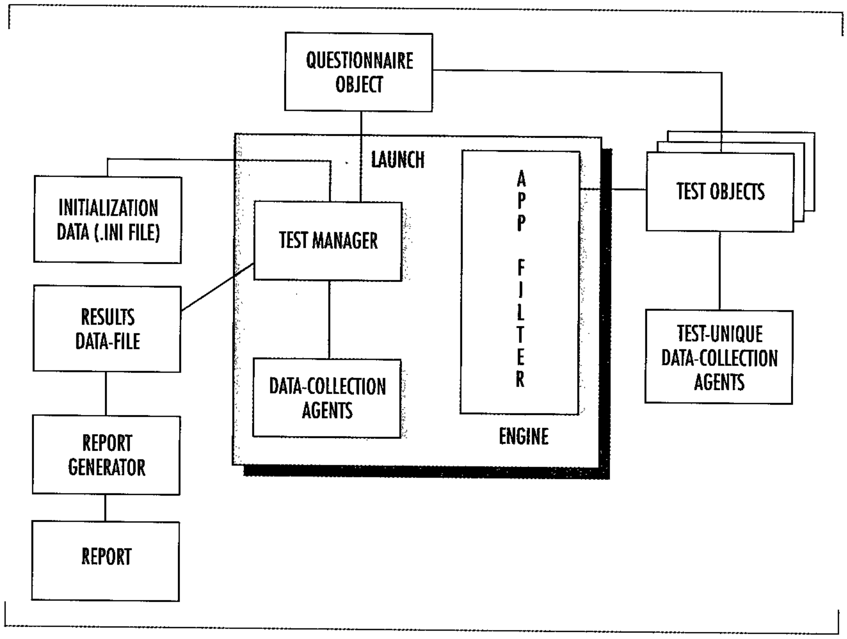


FIGURE 1. Block diagram of testing process.

is before the segment is installed; software developers provide this information. The second time is after installation to obtain information that the test engineer can determine easily, but that the automated software would not be able to determine (or be able to determine uniquely). The questionnaires present multiple-choice questions, including an "other" choice, where applicable. These objects read from formatted text data files and are therefore dynamic and easily modified.

The data-collection form and the menus presented are also dynamic. They are created by the compliance engine at runtime. The report generator merely specifies the state of each test object. This facilitates the additional tests as well as additional report formats. The testing process also tracks the test engineer's name and reports separately any tests that were waived or overridden. The test team found that the software to automate first should be that which provides the most efficient and largest payoff [2]. A phased approach to automation has proven most successful.

Designing the Automation Process

The requirements gathering described above also will yield one or more processes. In the beginning of software testing, all processes may not yet be in place. It is equally important to design the process. The automation software works within a process. This process will depend on the automation, which, in turn, will depend on the process recursively; often, both should be designed at the same time.

If testing has not been automated previously, the process will need a major rework. When automated testing is in place, personnel may be available to be tasked elsewhere. This will not be true in the beginning since the automation will also be undergoing testing. It is important to account for the testing assets that will be displaced by the automation. After some time, however, resource management will have to account for where to move these displaced testing assets.

Personnel Management Considerations

Test engineers, who will need to be trained how to use the automation, may have significant technical expertise. The main concern is to induce the test team to accept the new paradigm in which automation is replacing some of their expertise. It is not uncommon to see a reluctance to accept or attempts to discredit the automation.

In the author's experience, the best way to prevent this reluctance is to have the more experienced test engineers help with programming the algorithm-design phase if they are capable. It is important to get them to "take ownership" of the new paradigm. At least a small team of programmers will be required to help with the debugging and enhancing of the automation. The consistency and time savings will pay for these personnel.

The program manager also can redirect some personnel into a quality-assurance role to ensure that the output of the testing is of the required quality. Quality assurance is especially important if the results will be used outside of the test laboratory.

REVIEW

To achieve success in replacing the manual software testing process with an automated testing process, the test engineer must complete the following actions:

1. Document the current testing process.
2. Identify common processes.
3. Complete the following steps in parallel
   a. Design the automation software.
   b. Design the automation process.
   c. Encourage acceptance by the test engineers by inducing them to "take ownership" of the new process.
   d. If appropriate, consider using an object-design for the automation. Think about using a separate object for each test. Using objects helps when individual tests need to be modified.
   e. Think about a design that allows new tests and new reports to be added with few or no changes to the underlying data-collection (engine) process.
   f. When designing the software, think of the "big picture." How can this "machine" be used as part of a bigger system? Perhaps, instead of generating a report, the output could be used as input to a database. In this case, the report capabilities of the database could be used or the defects could be tracked automatically.

CONCLUSION

Replacing manual software testing with automated software testing can yield numerous rewards. A repeatable test process is the major advantage, leading to improved software quality and avoidance of a non-repeatable test. The depth of test coverage also can be increased, and the time requirements can be reduced. The combination of these two factors will improve the quality and cost savings of the software that supports DoD systems compliant with DII COE requirements. This testing methodology could be applied to testing software for government agencies outside DoD, such as the Department of Transportation Federal Aviation Administration and the Department of the Energy, both of which have exacting standards related to safety and security.

ACKNOWLEDGMENTS

This work was sponsored by the Defense Information Systems Agency–Defense Advanced Projects Research Agency (DISA–DARPA) Joint

Jack Chandler

BS in Computer Engineering, University of New Mexico, May 1991

Current Research: Automation of repetitive tasks and removal/reduction of subjectivity; collaboration research.

Program Office. The author thanks the entire test team, especially Pho Le, Steve Bitant, Will Greenway, Todd Webb, Randy Schiffman, MAJ Greg Csehoski and CAPT Stuart Kurkowski.

REFERENCES

1. Dustin, E. 1997. "Process of Introducing Automated Test Tools to a New Project Team," *Proceedings of the Rational User Conference.* URL: http://www.autotestco.com/html/sld001.htm
2. Pettichord, B. 1996. "Success with Test Automation," *Proceedings of Quality Week 96*, URL: http://www.io.com/~wazmo/succpap.htm

❖

Systems Integration Facility: Past, Present, and Future

David P. Andersen
SSC San Diego

Karen D. Thomas
Digital Wizards, Inc.

ABSTRACT

This paper traces the development of SSC San Diego's Systems Integration Facility (SIF) and the Combined Test Bed (CTB) that, together, provide a flexible, fully integrated multi-platform test capability used by dozens of multiservice and multinational testing organizations to ensure the interoperability of tactical data link systems. The paper describes unique PC-based Data Link Test Tools vital to Link-16 testing components. It also chronicles work of the major command, control, communications, computers, and intelligence (C^4I) interoperability testing organizations, such as Naval Sea Systems Command's Distributed Engineering Plant (DEP), and describes how the SIF/CTB will continue to support future tactical data link testing.

The Systems Integration Facility (SIF) opened in 1990 in Building 600 at SSC San Diego to support the Navy's first Joint Tactical Information Distribution System (JTIDS) developmental test program. Developed out of a need for a controlled, repeatable test environment to verify JTIDS terminal performance and combat systems interoperability, the SIF has become the Navy's leading laboratory for tactical data link interoperability testing.

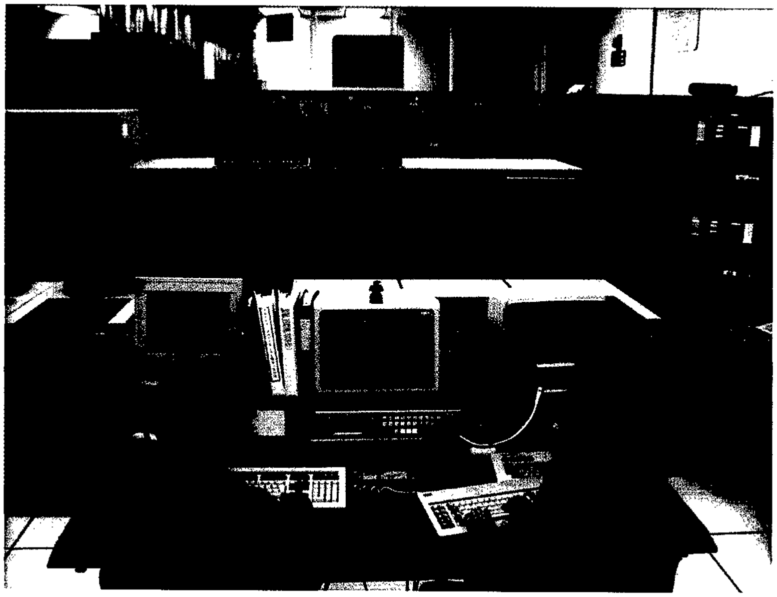
Sharing near-real-time tactical data in a distributed, interoperable, and secure environment is a critical segment of warfighter universal information access. Tactical data links, specifically Link-11 and Link-16, now the Department of Defense's primary data link, and the future Link-22, provide this capability to Navy, Joint, and Allied forces.

SSC San Diego has been involved with tactical data link development, test, evaluation, integration, and life-cycle support since the early 1960s. Under sponsorship of the Space and Naval Warfare Systems Command's Advanced Tactical Data Links Program Office (PMW 159), the SIF has played an integral role as the central node of a complex stimulation/simulation environment for land-based testing and evaluation of Link-16 components and systems and integration with other data link systems. SIF operations are part of SSC San Diego D45, the Tactical Systems Integration and Interoperability Division.

The first- and second-generation Link-16 uses the JTIDS data terminal, which provides multiple-access, high-capacity, jam-resistant digital data and secure voice communication, navigation, and identification information to various command and control and weapons host platforms. The JTIDS terminals encompass software, radio frequency (RF) equipment, and the waveform they generate. Link-16 requires JTIDS terminals and host combat systems such as the Advanced Combat Direction System (ACDS) or Aegis Command and Decision (C&D), processors such as the shipboard Command and Control Processor (C^2P) or the F14-D Mission Computer, Link-16 antennas, other hardware, software, and displays. The C^2P was developed at SSC San Diego to provide data forwarding and translation between Link-16, Link-11, and Link-4A.

Using Time Division Multiple Access communications architecture, Link-16 terminals transmit information in the Tactical Digital Information Link-Joint (TADIL-J) message format. A common communications net is thus provided to a large community of airborne and surface elements within line of sight, and the network can be extended to platforms beyond line of sight by using one or more members of the net, or any Link-16 terminal, as relays.

When the Navy JTIDS developmental test program was established to verify the technical adequacy of the JTIDS terminal and the integration of Link-16 into designated Navy host combat platforms, a land-based laboratory environment for terminal specification testing and multi-platform integration/interoperability testing was needed as a cost-effective precursor to live platform testing and to allow problem resolution.



The early SIF. Many of the original components of the Systems Integration Facility shown in this 1992 photo have since been replaced by "New SIF" distributed components hosted on personal computers, enabling an infinite number of system configurations that can be tailored to support a large number of test and training scenarios.

The SIF was designed to utilize a complex multi-computer simulator/stimulator connected to eight JTIDS terminals linked together by an RF network that introduced propagation delays and attenuation into tactical message traffic. The test bed also included external communications equipment, an antenna, test scripting, data storage, reduction, and analysis equipment, and various interfaces.

By September 1991, the SIF provided a fully integrated multi-platform functional testing capability. The SIF/Combined Test Bed (CTB) used intermediate processors to tie together the SIF and the Combat Direction System Development and Evaluation Site (CDES) laboratory in the same building, the E-2C Software Support Activity (SSA) laboratory in Building C-60, and the F-14D Mission Computer Subsystems Software Development Laboratory at Pt. Mugu, California. The CDES contains shipboard combat system configurations and programs for testing CV, LHD, Aegis CG/DDG, and LHA platforms. It also serves as the primary development and testing laboratory for the C^{2}P. The E-2C laboratory consists of actual E-2C Airborne Tactical Data System (ATDS) software and hardware components. The F-14D facility consists of actual F-14D software and hardware components.

By early 1993, the CTB was extended to the Aegis Combat Systems Center, Wallops Island, Virginia, for testing and integrating Aegis combat systems. Later, to support developmental testing of the new generation of Link-16 terminals, the Multifunctional Information Distribution System (MIDS), the F/A-18 Advanced Weapons Laboratory at the Naval Air Warfare Center, China Lake, California, was added to the CTB.

The SIF is the central node of the CTB, providing a central script controller to run the test information and direct it to real or simulated host systems that control the appropriate terminal type in the SIF terminal farm. The unique JTIDS RF simulation environment in the SIF provides connectivity between the SIF terminals, with digital propagation delays and attenuation matched to a scripted scenario. To support exercises that require live transmission rather than SIF RF network simulations, two

Shipboard combat system equipment in the Combat Direction System Development and Evaluation Site (CDES) laboratory in 1992. The CDES is an integral part of the Systems Integration Facility/Combined Test Bed for data link testing.

JTIDS antennas were installed on the roof of Building 600. Mobile JTIDS vans and portable JTIDS units, called mini-racks, were developed by the SIF team for deployment in other test locations or for installation on surface vessels. As tests were conducted, other SIF systems enabled collection of the test data for later replay and analysis.

The concept for JTIDS development and integration was to proceed through increasingly complex testing, from technical evaluation of the terminal to integration with the C^{2}P, then with the ACDS, and, finally, with air programs. Following the initial terminal testing program in the early 1990s, the SIF/CTB and D45 test and evaluation team members played major roles in the JTIDS and C^{2}P technical evaluation (TECHEVAL) processes that paved the way to a major milestone in the Link-16 program—the successful completion in 1994 of the required operational evaluation (OPEVAL) of the JTIDS and C^{2}P development program during the USS *Carl Vinson* (CVN 70) Battle Group's deployment to the Persian Gulf. This important step in the introduction of Link-16 and the C^{2}P into the Fleet was the culmination of years of development work by Navy activities and supporting contractors. It was also the beginning of new challenges for the SIF/CTB.

Early in the development of the SIF test bed, it was realized that significant modifications would be needed to support emerging test requirements. New capabilities were being added to Link-16 terminals. The new MIDS program was being planned, and the SIF would be the lead laboratory for terminal testing and integrating MIDS into Navy platforms, under sponsors PMW 159 and the MIDS International Program Office (PMW 101). The MIDS is a smaller, lighter weight terminal that maintains all JTIDS functionality. C^{4}I interoperability testing needed to expand to support Joint service and multinational interoperability scenarios, and interoperability testing needed to support operations in multilink environments.

While the SIF/CTB had provided valuable feedback to the Navy's JTIDS Program Office concerning the functional performance of Link-16 terminals and integration of the terminals with combat systems, its capability to support multi-service and multinational integration and interoperability testing was somewhat limited. A method of easily interfacing multi-service and multinational host combat systems was needed, as well as a system for addressing multi-link issues. These requirements led to the next developmental phase of the test bed.

Cost/benefit studies conducted by systems engineers from SSC San Diego and supporting contractors concluded that while the equipment in the

SIF was capable, it was costly to maintain and difficult to modify. The long-range functional requirements for the SIF/CTB could best be met by a complete re-engineering of the test bed's systems.

In the mid-1990s, development began on "New SIF" architecture. Its goal was to be a system with greater capability that could respond quickly and cost-effectively to the rapid evolution in functional requirements and could provide cost-effective test and evaluation support to any tactical platform, regardless of location or terminal availability. The "New SIF" would use commercial-off-the-shelf IBM-compatible personal computers and the OS/2 operating system that accommodated the robust multitasking required by the systems and provided a friendly graphical user interface. "New SIF" systems would be based on common software architecture to allow rapid development and flexibility when requirements changed.

By 1995, the Link-16 Gateway, now known as the Data Link Gateway (DLGW) system, was developed by D45 with contractor support to connect hosts at remote laboratories to the JTIDS and MIDS terminals in the SIF. The versatile PC-based Gateway system permitted the interfacing of multiple terminal farms, development laboratories, software support activities, live assets, and certification and simulation activities, forming a single extended Link-16 network for testing and integration. The Gateway system is composed of multiple DLGW units linked by secure dial-up phone lines or higher speed communications systems. Each DLGW can function as a host emulator, as a terminal emulator, or as a network monitor. The Gateway software provides a suite of functions that allows users to participate in data link exercises, and monitor, control, record, and analyze data from the exercises.

Other PC-based Data Link Test Tools were developed, including the following:

- Script Controller for executing test scripts on the SIF script network.
- Simulation Interface Units (SIUs) for translating scenario data in SIF format to the format and protocol needed by specific simulation systems.
- TADIL-J Host Simulator, a scenario-driven or real-time tactical data system emulator that creates realistic participants for testing and training.
- Link-16 Engine to support interconnection of non-Link-16-capable systems to the Gateway system.
- Script Generator for creating test scripts that pass events to various Data Link Test Tools for processing on a Link-16 network.
- Data Analysis and Reduction Tool (DART) for post-test analysis.

The original SIF systems were replaced by the "New SIF" distributed components hosted on PCs and communicating through Transmission Control Protocol/Internet Protocol (TCP/IP) on an Ethernet local area network (LAN). Because the new systems were interconnected to operate as a single distributed system, the re-engineered test bed offered an infinite number of system configurations that could be tailored to support a large number of test and training scenarios.

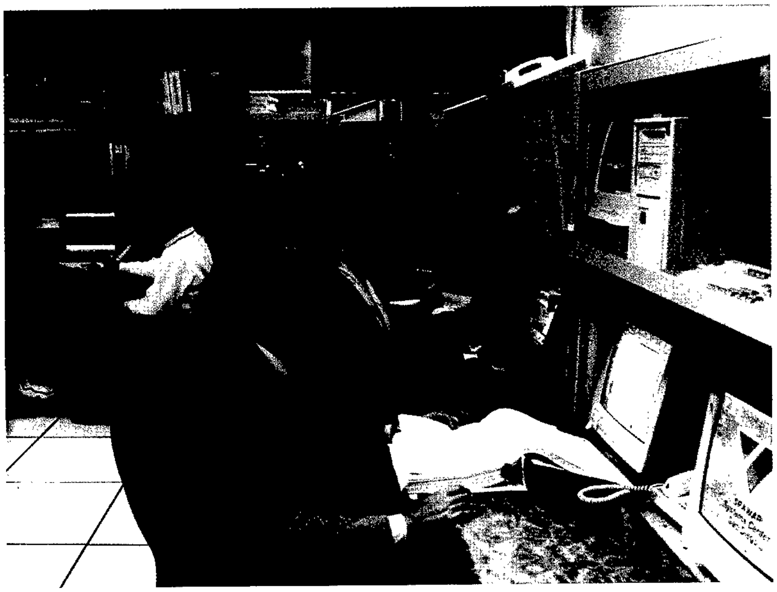
Each of the systems comprising the "New SIF" is a complex system in its own right, and each has evolved and continues to evolve to meet various new functional requirements. The SIF/CTB is a meta-system whose components are interconnected and mutually supporting. Within the SIF itself, systems communicate over the Script Net LAN. The remote sites are connected in a wide area network by the DLGW system, which multiplexes Link-16 and scenario data between sites. At the remote sites, SIUs convert the scenario data into the form needed by the site-specific simulation system so that all systems are not only communicating in the same link environment, but also participating in a single coordinated scenario.

By 1996, the "New SIF" began to evolve into a major hub for Joint and multinational C^{4}I interoperability testing and training, as well as a facility for testing new Link-16 terminal types such as the MIDS. Today, Data Link Test Tools provide Link-16 connectivity between the SIF and more than 100 Joint and international test and software support facilities, as well as all SSC San Diego C^{4}I laboratories, including the Research, Evaluation and Systems Analysis (RESA), the Reconfigurable Land-Based Test Site (RLBTS), and the Global Command and Control System (GCCS). By installing a DLGW system at each of the remote facilities and linking them by telephone lines or high-speed circuits, a Gateway network is created. This connectivity enables a worldwide TADIL and systems interoperability test capability.

In addition to the unique combination of assets in the SIF/CTB, key to the success of the TADIL testing programs is one of the most experienced and knowledgeable Link-16 engineering and test and evaluation (T&E) teams in the world. The D45 T&E team has supported at-sea testing and engineering programs since the early 1990s. The team's extensive hands-on engineering experience from early Navy JTIDS terminal testing to the complex interoperability test programs of today has provided a valuable resource for testing and integration programs.

Today, a wide variety of JTIDS and MIDS testing activities are offered by the SIF/CTB, including terminal functionality and specification testing, pre-installation testing and checkout, relative navigation performance evaluation, JTIDS terminal network load testing, TDS-to-TDS interoperability testing, multi-TADIL/multi-platform interoperability testing, TADIL network performance evaluation, TADIL trouble report testing, TADIL standards certification testing, new TADIL "proof-of-concept" analysis, and live fleet service support. In addition, the test bed supports production testing of JTIDS terminal firmware upgrades for the Command and Control Fleet Engineering Division's JTIDS/MIDS SSA (SSC San Diego D64), and Product Acceptance Testing (PAT) and Functional Interoperability Testing (FIT) for the C^{2}P SSA.

Scores of testing organizations have used the SIF/CTB resources. One of the first to use the DLGW systems and the "New SIF" for integration and interoperability was the Theater Missile



The SIF today. PC-hosted Data Link Test Tools communicating via TCP/IP protocols on an Ethernet local area network have replaced the original systems in the SIF, which is now the Navy's leading laboratory for tactical data link interoperability testing.

The team today. The D45 government and contractor teams for SIF/CTB operations, Data Link Test Tools development and support, and data link testing and evaluation.

Defense System Exerciser (TMDSE), a program of the Ballistic Missile Defense Organization (BMDO) to integrate the entire TMD family of systems and test interoperability issues between the various TMD systems. The SIF/CTB supports certification testing programs conducted by the Joint Interoperability Test (JIT) network directed by the Joint Interoperability Test Command (JITC) and by the Navy Center for Tactical Systems Interoperability (NCTSI). The SIF is the lead laboratory for the ongoing MIDS Low Volume Terminal (MIDS-LVT) and MIDS on Ship (MOS) test and evaluation programs. The SIF/CTB and the D45 T&E teams have played a significant role in the North Atlantic Treaty Organization (NATO) program to test the Standard Interface for Multiple Platform Link Evaluation (SIMPLE), and the test bed has been used in many of the Navy's Commander, Operational Test and Evaluation Force (COMOPTEVFOR) testing programs. The SIF has been accredited by COMOPTEVFOR for operational testing of the rehosted C^{2}P and the MOS terminal.

The SIF/CTB is a development and land-based testing and evaluation environment for the C^{2}P, the rehosted C^{2}P, the Common Data Link Management System (CDLMS) and, most recently, for the Multi-TADIL Capability (MTC) Global Command and Control System–Maritime (GCCS–M) program. The MTC will provide a standard and interoperable data link capability for exchanging information on TADIL-A, TADIL-B, TADIL-J, and Satellite TADIL-J (S-TADIL-J) across the entire Joint environment. The SIF has been equipped with computer systems dedicated to the MTC program, currently developed for use with GCCS–M. To support interoperability and integration testing of Aegis ship classes and related subsystems, the Integrated Combat System Test Facility (ICSTF), a field activity of Naval Sea Systems Command (NAVSEA) located at SSC San Diego, has located its Aegis 5.3.7 test bed in the CDES.

The SIF/CTB has supported the Cooperative Engagement Capability (CEC) "Road to OPEVAL" integrated testing program since 1996. The world's most technically advanced air defense system, the CEC has been a top priority for the Navy to achieve its vision of network-centric warfare, and has involved many systems interoperability issues during its development. Support is also provided for the JTIDS Range Extension (JRE) program, which involves transferring Link-16 live satellite and S-TADIL J transmissions through the C^{2}P. S-TADIL-J was developed by the Navy to provide Link-16 connectivity when that connectivity is lost or affected by range limitations.

The SIF/CTB and Data Link Test Tools have become integral components of the extensive land-based Battle Group test bed of the Distributed Engineering Plant (DEP), a NAVSEA program designed to improve fleet readiness by identifying and resolving interoperability issues before deployments. The DEP connects, in real-time, land-based combat and battle management systems located in various Navy testing facilities across the U.S. The SIF is now home to the DEP's TADIL Operations Center (TOC). Information is exchanged among battle groups through Link-16 and Link-11, and DLGW Terminal Emulators located at each DEP Link-16 host site provide the Link-16 message exchange capability for the test bed. D45 provides the DEP TADIL team leaders.

To support government qualification testing of MIDS-LVT (PMW 101) production terminals, an environmental testing chamber is being installed in the SIF. First Article Qualification Testing (FAQT) of MIDS-LVT vendor terminals will include functional performance, interchangeability, and terminal compatibility tests. Following successful FAQT testing, the vendors will be allowed to competitively bid on full-rate production of the MIDS-LVT.

Today, more than 100 operational, test, training, and development sites around the world use the unique combination of interconnected Link-16 terminals, operational hardware and software, Data Link Test Tools, simulation systems, ship and air laboratory connectivity, live transmit/receive facilities, robust Link-11 capability, and the SIF's engineering, evaluation, and integration expertise to assist in the development and operational evaluation of tactical data systems.

Once begun as a single centralized JTIDS test bed with three remote development and test sites, the SIF/CTB has now become a powerful distributed network providing comprehensive operational testing and training support to the C^{4}I community worldwide. As additional data links are developed and as interoperability programs are expanded and new programs begin, this unique test bed is well prepared to accommodate the future needs of the Navy, Joint, and Allied nation testing communities it serves.

❖

David P. Andersen

BS in Mathematics, San Diego State University, 1985

Current Work: Link-16 (Joint Tactical Information Distribution System [JTIDS]/Multifunction Information Distribution System [MIDS]) Test and Evaluation Business Area Manager.

Karen D. Thomas

BA in Journalism, San Diego State University

Current Work: Principal Analyst/ Writer; computer systems engineering.

4

Simulation and Human–Systems Technologies ■

4 SIMULATION AND HUMAN–SYSTEMS TECHNOLOGIES

Advanced Distributed Simulation: Decade in Review and Future Challenges

Douglas R. Hardy and Elaine C. Allen, SSC San Diego

Kevin P. Adams, Charles B. Peters, and Larry J. Peterson
SSC San Diego
Michael A. Cannon, VisiCom
Jeffrey S. Steinman, RAM Labs
Bruce W. Walter, Greystone Technology, Inc.

ABSTRACT

As networking technologies and computer hardware performance advanced in the late 1980s, distributed simulation became a feasible way to provide military training at distant, sometimes remote locations. Efforts were made to advance the technologies surrounding distributed simulation, from networking protocols to the representation of the battlespace and its entities. The following SSC San Diego efforts supported advances in distributed-simulation-related areas throughout the 1990s and continue to support the next generation of 21st century simulation systems.

INTRODUCTION

The 1990s saw SSC San Diego continue to be the leader in Advanced Distributed Simulation (ADS) technologies for the U.S. Navy. SSC San Diego simulations supported worldwide users in training, assessment, analysis, testing, experimentation, and technology research. SSC San Diego supported network-centric simulations and joint-service objectives. The Center defined and advanced two major simulation protocol threads: the Distributed Interactive Simulation (DIS) protocol and the Aggregate-Level Simulation Protocol (ALSP). These protocols were the genesis of the latest and current Defense Modeling and Simulation Office's (DMSO's) standard: the High-Level Architecture (HLA) Run-Time Infrastructure (RTI).

SSC San Diego's simulation efforts supported a variety of venues that tested and experimented with the protocols over large distributed networks, and developed capabilities that supported the trend from service-specific to joint-service exercises. The major advanced distributed simulation efforts during the decade were provided by the following SSC San Diego simulation systems: the Research, Evaluation, and System Analysis (RESA) Simulation; the Marine Corps' Marine Air Ground Task Force (MAGTF) Tactical Warfare Simulation (MTWS); the Synthetic Theater of War (STOW) Advanced Concepts Technology Demonstration (ACTD); and the Joint Simulation System–Maritime (JSIMS–M). These simulations supported venues that included the construction of Joint Federation training exercises supported by RESA and MTWS through their development of ALSP interfaces. The support included the advent of ACTDs, with STOW emerging as the first ACTD, and further support was provided to a variety of subsequent ACTDs (e.g., Joint Countermine Operational Simulation (JCOS), Extending the Littoral Battlespace (ELB), and Joint Medical Operations–Telemedicine (JMO–T)) using DIS and eventually RTI protocols. Additional support has continued through experimentation in Fleet Battle Experiments (FBEs) and Joint Experimentation (JE) events. SSC San Diego simulations will continue to support these venues by improving existing simulations and by developing next-generation advanced distributed simulation systems that support joint-service operations, such as JSIMS–M.

The following section will briefly describe support provided by these SSC San Diego simulations, including some specific events, followed by the final section on future potential.

ADVANCED DISTRIBUTED SIMULATION (1990s)

As networking technologies and computer hardware performance advanced in the late 1980s, distributed simulation became a feasible way to provide military training at distant, sometimes remote locations. Efforts were made to advance the technologies surrounding distributed simulation, from networking protocols to the representation of the battlespace and its entities. The following SSC San Diego efforts supported advances in distributed-simulation-related areas throughout the 1990s and continue to support the next generation of 21st century simulation systems.

Research, Evaluation, and System Analysis (RESA)

The RESA simulation system has a 23-year history and has evolved to meet the Navy's ever-expanding needs for a constructive simulation system that focuses on theater-level naval operations. The capabilities of RESA to realistically simulate the naval warfare environment, generate streams of realistic scenario-driven data to C^4I support systems, and to interface with other models/analysis tools have led to its application in a wide variety of projects.

Throughout the 1990s and continuing today, the RESA system has provided the Navy with a stand-alone system to support a wide variety of applications, including systems analysis, concept of operations development, advanced technology assessment, and C^4I system simulation. In the early 1990s, the reliability of the system and its flexibility in adapting to the Navy's changing needs, led to its evolution into today's RESA system, fulfilling dual missions in the areas of joint-forces training and joint and naval research, development, test, and evaluation (RDT&E).

To fulfill the Navy's need for a naval training system within the U.S. Joint Forces Command (JFCOM) Joint Training Confederation (JTC), the RESA team aided in the design of the ALSP. Developed specifically for the JTC, the ALSP interface allowed the sharing of simulation information with other service constructive simulations including the Army's Corps Battle Simulation (CBS) and the Air Force's Air Warfare Simulation System (AWSIMS). Today, the ALSP JTC integrates a wide variety of models and simulations supporting joint forces and allied training at the command level, worldwide, in exercises such as Unified Endeavor at the U.S. Atlantic Command (USACOM) and Ulchi Focus Lens at the Combined Forces command in South Korea. In the mid-1990s, the Marine Corps MAGTF MTWS system, developed and supported by SSC San Diego, was integrated into the JTC, thus completing the inclusion of all joint-service warfare areas.

Concurrent with providing the Navy's system in the JTC, SSC San Diego was selected to participate in the design and development of the DIS protocols for the integration of joint-service constructive simulations, virtual models, and live-range entities. This task was accomplished in support of joint-service assessment, analysis, testing, experimentation, and technology research. The RESA system became one of the Navy's first DIS-compliant simulations, and it has been used in a variety of joint-service and allied studies sponsored by DMSO, the Defense Advanced Research Projects

Agency (DARPA), the Ballistic Missile Defense Office (BMDO), the Office of the Secretary of Defense (OSD), and the Office of the Chief of Naval Operations (OPNAV). As the naval component in joint-service distributed projects, the RESA system has contributed to developing and testing command and control structures, operational plans, concepts of operation, and analyses. Areas of study include analyses of the Cooperative Engagement Capability (CEC), the next-generation aircraft carrier (CVNX), the Zumwalt-class 21st century destroyer (DD 21), and Joint Theater Missile Defense Attack Operations.

The extensive simulation capabilities of RESA, coupled with its record of reliable operations and transportability, have not only resulted in its use at a number of facilities for a variety of applications but have also led to its use in providing the core simulation infrastructure for other simulation developments such as the CounterMeasures Analysis Simulator (CMAS), Space and Electronic Warfare Simulation (SEWSIM), the Air Warfare Simulation System (AWSIMS), and the Battleforce Electro-Magnetic Imagery (EMI) Evaluation System (BEES).

The history of the RESA system not only lends merit to SSC San Diego's current reputation as a prominent leader in the design and development of distributed simulation systems, but also attests to SSC San Diego's status as a true pioneer in the world of modeling and simulation (M&S).

Marine Air Ground Task Force (MAGTF) Tactical Warfare Simulation (MTWS)

The MAGTF MTWS system, developed and supported by SSC San Diego, is a constructive simulation that provides exercise control services and tactical combat simulation capabilities to support tactical training exercises. Development of MTWS began in 1989. In 1995, the system was formally accepted by the Marine Corps as the replacement for the Tactical Warfare Simulation, Evaluation, and Analysis System (TWSEAS). MTWS supports all aspects of MAGTF combat operations, including air, ground, maritime, and amphibious operations, in a multisided environment to permit creation of the widest possible range of tactical conditions to challenge staff decision-making. The MTWS Analysis Review System (MARS) component provides the training audience with exercise review, analysis, and replay capabilities.

In the mid-1990s, the MAGTF MTWS system was integrated into the JTC via an ALSP interface. The MTWS ALSP interface supports a wide variety of air, ground, and surface interactions with other ALSP confederates. In a confederation with multiple MTWS actors, the interface supports ground-to-ground interactions; this is unique within the ALSP confederation. Besides the ALSP interface for supporting interoperation with the JTC, a DIS interface was developed to support real-time simulation interoperability with other DIS simulations, such as the Joint Semi-Automated Forces (JSAF) simulation. MTWS was used in conjunction with JSAF to support modeling and simulation for the ELB ACTD in 1999. MTWS also interfaces to C4I systems such as the Global Command and Control System (GCCS), providing scenario-based track update information via over-the-horizon (OTH)-GOLD messages, and a variety of Intel-related U.S. Message Text Format (USMTF) messages.

In its original configuration, MTWS operated as a set of simulation applications distributed across a networked suite of TAC-4 HP processors, connected via a central hub to a second network of TAC-3/4 user stations.

The simulation applications—ground combat, air operations, ship-to-shore, logistics, etc.—can be distributed over as many as six host processors, or all can run on a single host processor, at the user's option, depending on the size, scope, and intensity of the scenario. The user stations provide a tactical map display supporting both vector and raster map images, as well as various exercise definition, control, and reporting functions. In early 2001, the TAC-3/4 user stations were replaced by PC/Win32 workstations, which provide enhanced functionality with increased performance.

As the TAC-4 hardware is phased out, and the functionality and capacity of the system continue to increase, MTWS is evaluating the benefits of migrating the remainder of the system to another platform(s). This includes migration to more platform-independent development tools (e.g., compiler, etc.). Also, MTWS expects to introduce a Web-based After-Action Review (AAR) system this year, which will significantly enhance the potential to support remote training.

Synthetic Theater of War (STOW)/Joint Semi-Automated Forces (JSAF)

STOW, developed in the mid-1990s, was based on the culmination of several advanced research projects sponsored by DARPA in the early 1990s. These projects spearheaded efforts to advance technologies for the next generation of computer-generated forces and distributed simulation; specifically in areas of aggregation/deaggregation, high vs. engineering fidelity, scalability (handling large numbers of distributed objects), and DIS protocols. STOW Europe (STOW-E) exploited these technologies by integrating constructive, virtual, and live simulation in a major joint exercise in 1994 called Unified Endeavor (Reforger). The exercise was held primarily in Germany but was distributed to sites in England and the U.S. In 1995, STOW transitioned to an ACTD.

STOW evolved from Simulation Networking (SIMNET) protocols to DIS protocols to DMSO's standard HLA RTI protocols. In 1997, STOW became the largest federation ever to use the newly mandated HLA RTI protocols. The main product that the STOW ACTD transitioned was a joint distributed simulation capability called Joint Semi-Automated Forces (JSAF).

Currently, JSAF primarily supports the JE events for JFCOM at the Joint Training and Analysis Simulation Center (JTASC) in Suffolk, VA. The JFCOM Experimentation Directorate, J9, is now the operational sponsor and makes extensive use of JSAF for Human-in-the-Loop (HITL), virtual experiments. The U.S. Navy's Maritime Battle Center uses JSAF as the core simulation for its Fleet Battle Experiments (FBEs). JSAF is also supporting the Joint Medical Operations–Telemedicine ACTD. A JSAF User's Group has recently been created to represent a broadening group of agencies making use of HLA-compliant JSAF technologies.

Joint Experimentation Using JSAF

Joint Experimentation 9901 (JE9901)

The JE9901 Experiment explored new approaches to JE in the context of investigating how future systems, especially sensor systems, can be used to defeat critical mobile targets in the form of theater ballistic missiles before they are fired. The Critical Mobile Target Cell (CMTC) was used to provide real-time tasking authority for the sensors, and then Automated Target Recognition was used to continuously track targets.

Attack Operations 2000 (AO 00)

The AO 00 Experiment used a war-game scenario with Sensor and Shooter Concept of Operations (CONOPS) for the 2007 timeframe. The experiment dealt with HITL acting as a black-box surrogate, deciding on which platforms and weapons systems/munitions were to be used in a synthetic environment. The AAR system logged the Experiment/Simulation in real-time for playback, thread analysis, and battle damage assessment. (See Figure 1 for sample synthetic environment.)

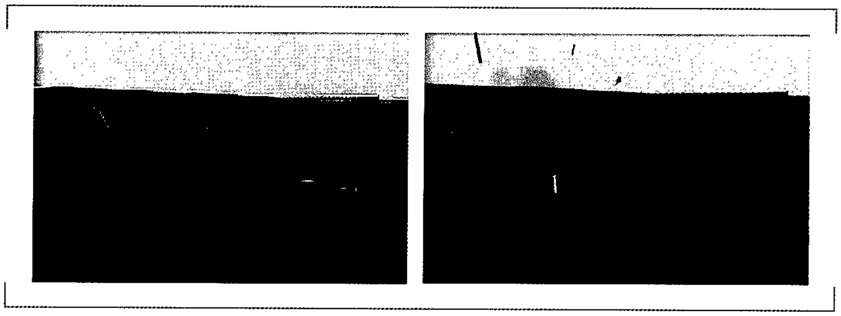


FIGURE 1. Realistic environments and dynamic terrain features have become a reality in simulation. A simulated vehicle crosses a bridge (left), and then the bridge is bombed and destroyed (right), making the bridge impassable by other vehicles. Bridging assets now exist that could build a simulated temporary bridge for forging the river.

ACTD Using JSAF

Joint Countermine Operational Simulation (JCOS)

The objective of the Joint Countermine (JCM) ACTD was to demonstrate the capability to conduct seamless mine countermeasure (MCM) operations from sea to land. The ultimate goal was to develop improved MCM equipment, operational concepts, and doctrine to support amphibious and other operations involving Operational Maneuver from the Sea (OMFTS), and to support the follow-on land operations.

Modeling and simulation played a key role in the JCM ACTD. JCOS was used to evaluate the operational use of countermine systems, to evaluate plans developed to accomplish exercise objectives, and to evaluate doctrine and tactics in a variety of scenarios and tactical situations.

The JCOS goal was to provide an end-to-end simulation capability for joint MCM operations. JCOS used and leveraged existing Advanced Distributed Simulation JSAF capabilities to meet this goal. With this approach, JCOS was able to simulate and rehearse joint warfighting operations in a mined environment across the operational continuum from deep water, through littoral, to inland objectives.

JCOS was used during the planning phases of two amphibious assault exercises that required extensive MCM operations. JCOS was also used during exercises to simulate a much more robust MCM component.

Joint Medical Semi-Automated Forces (JMedSAF)

The objective of the Joint Medical Operations–Telemedicine (JMO–T) ACTD was to provide a near-term capability to defeat time, distance, and organizational obstacles to effective Joint Health Service Support in austere and nonlinear operational environments.

The plan developed by SSC San Diego to provide M&S support for the ACTD was similar to that followed for the JCM ACTD, in which JSAF capabilities were enhanced in the specific domain area required. A comprehensive representation of Army, Air Force, Marine, and Navy medical treatment behaviors was developed to provide medical mission planning and rehearsal at a Joint Task Force/Commander in Chief (CINC) level that would be on a par with those employed by the combat branches. (See Figure 2.)

Specific capabilities developed include:

- Medical entities: hospital ships, medical treatment facilities, ambulances, helicopters, and individuals capable of being wounded or sick.
- Medical behaviors: combat injuries based on weapon/casualty type pairings and defined medical patient codes, disease and nonbattle injuries determined on percentage of population at risk, medical facilities with staff, equipment, holding capacities, and evacuation assets.
- Medical C^2 reporting: a medical C^2 message interface to the Medical Equipment Workstation (MEWS) that will provide Annex Q (medical reports section of an OP Order) reporting as well as information on individual patient encounters.

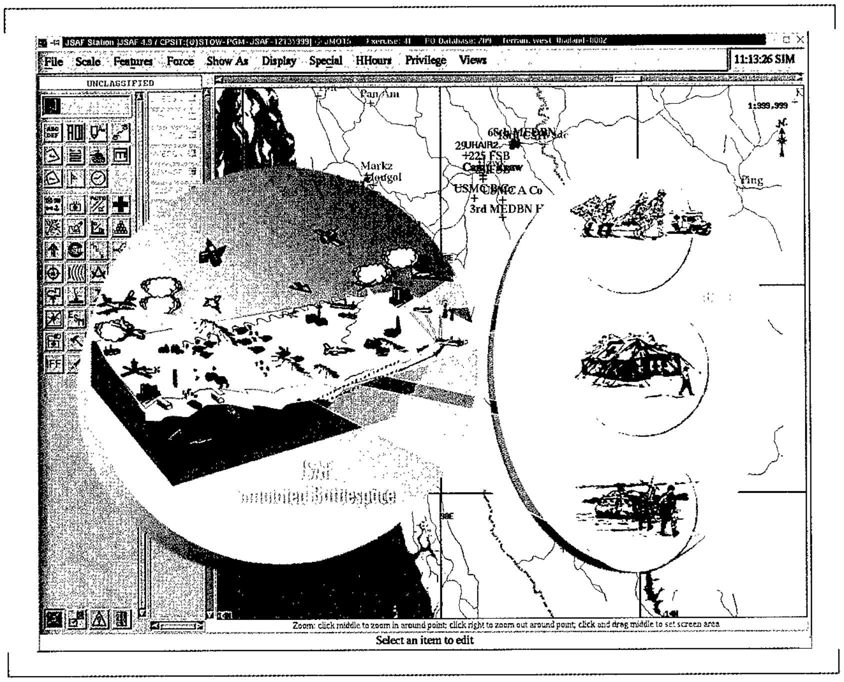

JSAF Station |JSAF 4.9 / CPSIT:(U)STOW-PGM-JSAF-12131999| JMOT5 Exercise: 41 PO Database: 709 Terrain: west thailand-0002
File Scale Features Force Show As Display Special HHours Privilege Views
11:13:26 SIM
UNCLASSIFIED
Pan Am
Markz Mongol
29UHAIR2
+225 FSB
USMC BMC A Co
3rd MEDBN F
1:999,999
Ping
Zoom: click middle to zoom in around point; click right to zoom out around point; click and drag middle to set screen area
Select an item to edit

FIGURE 2. JMedSAF is a medical extension of JSAF, providing the ability to simulate medical play in the simulated tactical battlespace. Medical play includes combat injuries, disease-related illnesses, and nonbattle injuries; medical treatment facilities, their staffs, and supplies; the evacuation of injured or sick, or subsequent return-to-duty; and interfaces to medical C^2 workstations.

JMedSAF has been demonstrated at Kernel Blitz '99 in conjunction with ELB ACTD (April 1999), in the Pacific Warrior Exercise CPX (November 1999), and in Cobra Gold 2000 (May 2000). Participation in Cobra Gold 2001 is also planned. JMedSAF will also be used (in conjunction with a distributed simulation from the Army's Training and Doctrine Command) to assess the effects of varying levels of medical support for future Objective Forces.

Extending the Littoral Battlefield

The main objectives of the ELB ACTD were to (1) expand battlespace connectivity in the littoral regions by using wireless network technologies and hand-held computing devices, and (2) further flatten the command and control structure for executing missions in austere and nonlinear operational environments.

The plan developed by SSC San Diego provided M&S support to the ELB ACTD Major System Demonstration (MSD) #1 in order to (a) accomplish greater realism for the common tactical picture, (b) enhance situational awareness of the battlespace, (c) increase the density of message traffic to C^4I systems, and (d) provide a mechanism to support testing events when limited resources were available. The simulation objectives were to:

- "Round out the battlespace" by using simulated entities as required for testing and demonstration (e.g., Supplemental Blue and Opposing Force Units, Ships, End User Terminals [EUTs], P3C, etc.)
- Provide certain simulated sensor message feeds (e.g., Joint Surveillance Target Attack Radar System [JSTARS], Tactical Remote Sensor System [TRSS], Guardrail, and unmanned aerial vehicle [UAV])

- Stimulate the ELB Watch Officer Workstation with OTGold and USMTF messages
- Stimulate the RMTP network with simulated EUT message traffic (JUnit and SALUTE POSREPs)

ELB employed two war-gaming simulation systems to accomplish these objectives: MTWS and JSAF. The simulations used their specialized strengths to provide the required functionality. JSAF was primarily used for higher fidelity amphibious, mine, and special operations, while MTWS was primarily used for its higher echelon battlespace representation, including rear area force and other massed troops with fewer computing resources necessary.

Joint Simulation System–Maritime (JSIMS–M)

JSIMS–M began development in the late 1990s and promises to be the next generation of advanced distributed simulation. JSIMS–M is being developed as a state-of-the-art simulation system in conjunction with the overall JSIMS Alliance. The development environment is based on object-oriented principles that use automated-code generation tools for overall reduced costs in the development and maintenance phases. In 2000, JSIMS–M became responsible for developing the Simulation Engine for the JSIMS Alliance. The Simulation Engine is based on a Government off-the-Shelf (GOTS) parallel discrete event simulation called Synchronous Parallel Environment for Emulation and Discrete Event Simulation (SPEEDES). This high-tech simulation can support faster-than-real-time operations, multi-processor systems, and simulation repeatability. SPEEDES is a simulation framework that supports simulation interoperability across a variety of parallel and distributed platforms (see Figure 3.)

SPEEDES development was initiated in 1990 by the National Aeronautics and Space Administration (NASA) at the Jet Propulsion Laboratory and was one of a number of simulation infrastructure projects initiated in the early 1990s that explored simulation interoperability over different computing platforms. The primary goal of SPEEDES was to provide interoperability between objects distributed across large numbers of processors while using a common simulation engine. A key feature of SPEEDES is its ability to preserve causally correct event processing in a repeatable manner without sacrificing parallel performance or constraining object interaction.

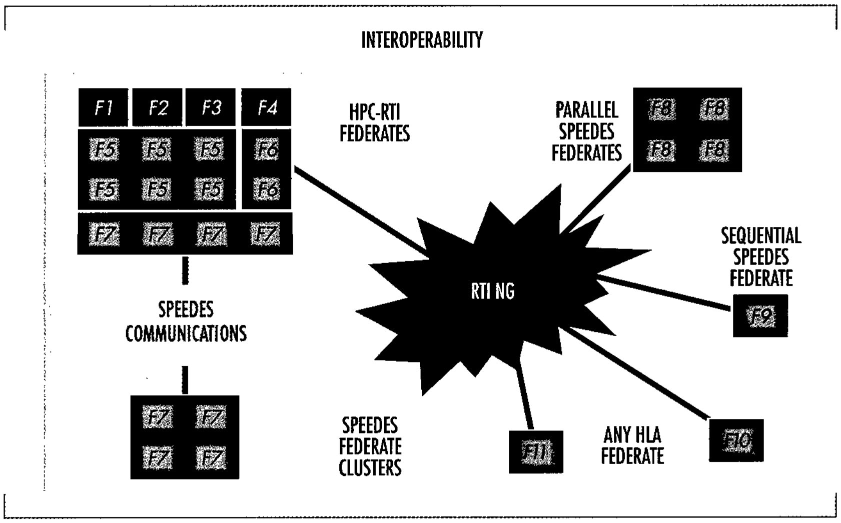


FIGURE 3. SPEEDES is a parallel discrete event simulation engine. The flexibility of the SPEEDES environment is depicted above and provides the capability of executing on one or many processors. The interoperability is maintained between SPEEDES nodes and any other HLA RTI federates.

Currently, several Department of Defense simulation projects use SPEEDES to provide all or part of their core infrastructure. Besides JSIMS, there is the Joint Modeling and Simulation System (JMASS), the Extended Air Defense Test Bed (EADTB), the Joint National Test

Facility's (JNTF's) Wargame 2000, the High Performance Computing and Modernization Office (HPCMO) infrastructure, and the Defense Modeling and Simulation Office (DMSO) project in support of a Human Behavioral Representation Test Bed.

The JSIMS program will provide a simulation environment capable of meeting a broad set of requirements for training and mission rehearsal. JSIMS is a single, distributed, seamlessly integrated modeling and simulation environment. The system provides the software and hardware infrastructure necessary to support multiple training, planning-and-analysis rehearsal, education, and doctrine development events in a variety of composable configurations. JSIMS–M is the component of JSIMS necessary to satisfy Navy training needs. JSIMS–M provides the capability to JSIMS to represent all aspects and elements of the maritime operational environment needed to support the execution of joint and service scenarios, and to train JTF and JTF component staffs. JSIMS–M will ultimately replace RESA and the Enhanced Naval Wargaming System (ENWGS) in joint and Navy training environments.

The overall development of JSIMS is the responsibility of an executive structure called the JSIMS Alliance, which relies on software development from multiple Domains. The Domain Agent (DA) for the maritime component of JSIMS is the Space and Naval Warfare Systems Command (SPAWAR) (PMW 153).

As a result of Alliance-wide reorganization occurring at the end of 1999, the Maritime domain has been identified as the development domain responsible for both Maritime objects in simulation (such as naval vessels, weapons), the ocean acoustic propagation loss data, and the development of certain common components that will provide data and software services to all components of the Joint Simulation. JSIMS Maritime common components' products include the Common Components Simulation Engine (CCSE), the Common Algorithms Support Services (CASS), and the Model Driver Database Diagnostic Interface (MDDI).

JSIMS is a multi-domain, cross-service military simulation system built on HLA. The HLA enables simulation objects modeled in multiple domains to be brought together into an application-specific joint simulation known as a federation. Within the HLA Architecture, multiple Simulation Object Models (SOMs) and supporting libraries can be accessed to provide various objects and services to compose a Federation Object Model (FOM). The coordination of object models is achieved through an RTI, which operates at the federate level. Services to the RTI are provided through the CCSE directly or provided through a specialized interface, depending on the architecture of the participating federate.

High Performance Computing (HPC)

High Performance Computing (HPC) initiatives were supported throughout the decade and have focused on the ability to use distributed, extremely high performance parallel-processing systems. SSC San Diego receives funding from the HPC Modernization Program (HPCMP) through its Common HPC Software Support Initiative (CHSSI). The CHSSI Force Modeling and Simulation (FMS)/C4I FMS Computational Technology Area supports the development of a simulation run-time infrastructure for HPC (HPC-RTI). Its immediate purpose is to greatly enhance computing capabilities for HLA distributed simulations. The HPC-RTI allows parallel computers to manage multiple HLA federates on a single machine while partaking in a distributed HLA simulation (Federation). The engine for the HPC-RTI is SPEEDES, which provides time management, data distribution management, object management, etc. SPEEDES, is currently the simulation engine for JSIMS and the BMDO Wargame 2000 system, and is currently being integrated into JMASS. The HPC-RTI then provides an HLA structure for SPEEDES.

ADVANCED DISTRIBUTED SIMULATION 2000+

As we move forward into the 21st century, JSAF and HPC will continue to support advanced distributed simulation efforts, and JSIMS–M will become part of the next generation of simulations.

JSAF will continue JE support with Unified Vision 2001 (UV01), Millennium Challenge 2002, and Olympic Challenge 2004. The mission of UV01 is to support the JFCOM Campaign Plan 2001. The Joint Experimentation Directorate (J9) is conducting a concept refinement experiment integrating Rapid Decisive Operations and its supporting functional concepts, as well as preparing for Millennium Challenge 2002 and Olympic Challenge 2004.

The HPC-RTI goal is to integrate into the GCCS as part of the Defense Information Infrastructure Common Operating Environment (DII COE) in order to provide a modeling and simulation capability to the Warrior in support of C^4I. HPC will also investigate further enhancements to SPEEDES, including an integration of a Common Reasoning Engine (CORE) along with other behavior-capture mechanisms and near-optimal decision-making mechanisms to provide commander objects in a distributed parallel environment through the HPC-RTI. The HPC-RTI will provide scalability of simulation size (large numbers of objects, large numbers of decision mechanisms, and large numbers of human-like behaviors) and reliable performance with real and faster-than-real time.

JSIMS–M will continue to investigate performance enhancements to SPEEDES and critical functionality improvements. The fundamental challenge for this parallel discrete-event simulation is to efficiently process events concurrently on multiple processors while preserving the overall causality of the system as it advances in simulated time. While JSIMS–M is currently being developed as the next generation of advanced distributed simulation based largely on the simulation engine (SPEEDES) and its future direction, the generation-after-next should also evolve with the advance of simulation technologies. Enhancements in performance and affordability of parallel systems, and automation of development and interface frameworks will lead to robust, high-speed, quickly reconfigured simulations to support a plethora of military and commercial uses. The simulations will cross domains from training, to analysis, to concept exploration, to test and evaluation and more. The simulation will simplify the support of training venues that include training at multiple echelons simultaneously. For example, the medic will be trained in triage or on a patient simulator by using virtual simulators, while another medic is in the field in a live exercise entering patient encounters using a Palm-top. The encounters are fed into the overall simulation and provide medical situational awareness to the medical commander and his staff. While using the same simulation, the staff will be able to take the "real" C^4I picture off-line, and run faster-than-real-time to evaluate and analyze various courses of action. These courses of action will be interactive and allow different inputs and constraints to be imposed. The ability to accomplish most of this exists, but the ability to do it with ease, and at reasonable cost, is still difficult.

Some challenges still facing future simulation include:

Scalability/Adaptability: Can a simulation be effectively tailored to support the task at hand both in size (footprint) and functionality? For instance, can the simulation be run on a laptop to train an individual or small group while in transit to an operational area? Can it be scaled to support large task forces over multiple operational areas, including coalition forces?

Network Capacities/Load Balancing: Can the simulation be distributed via various network capacities to the sites and/or platforms involved? For instance, can the simulation be used over limited bandwidth connections to a platform or perhaps limited because of security requirements? Do nodes on multiprocessing platforms have the approximate same workload?

Multi-Echelon Training: Can modeling and simulation be cost-effective for supporting integrations of constructive, virtual, and live simulations? Can these integrated simulation solutions support multi-echelon training at the appropriate fidelity for each echelon? Can interface frameworks be developed that make interoperability between these domains affordable?

Multiple Domains: Can a single simulation architecture have the flexibility to extend through domain areas (training, analysis, research, experimentation, etc.)?

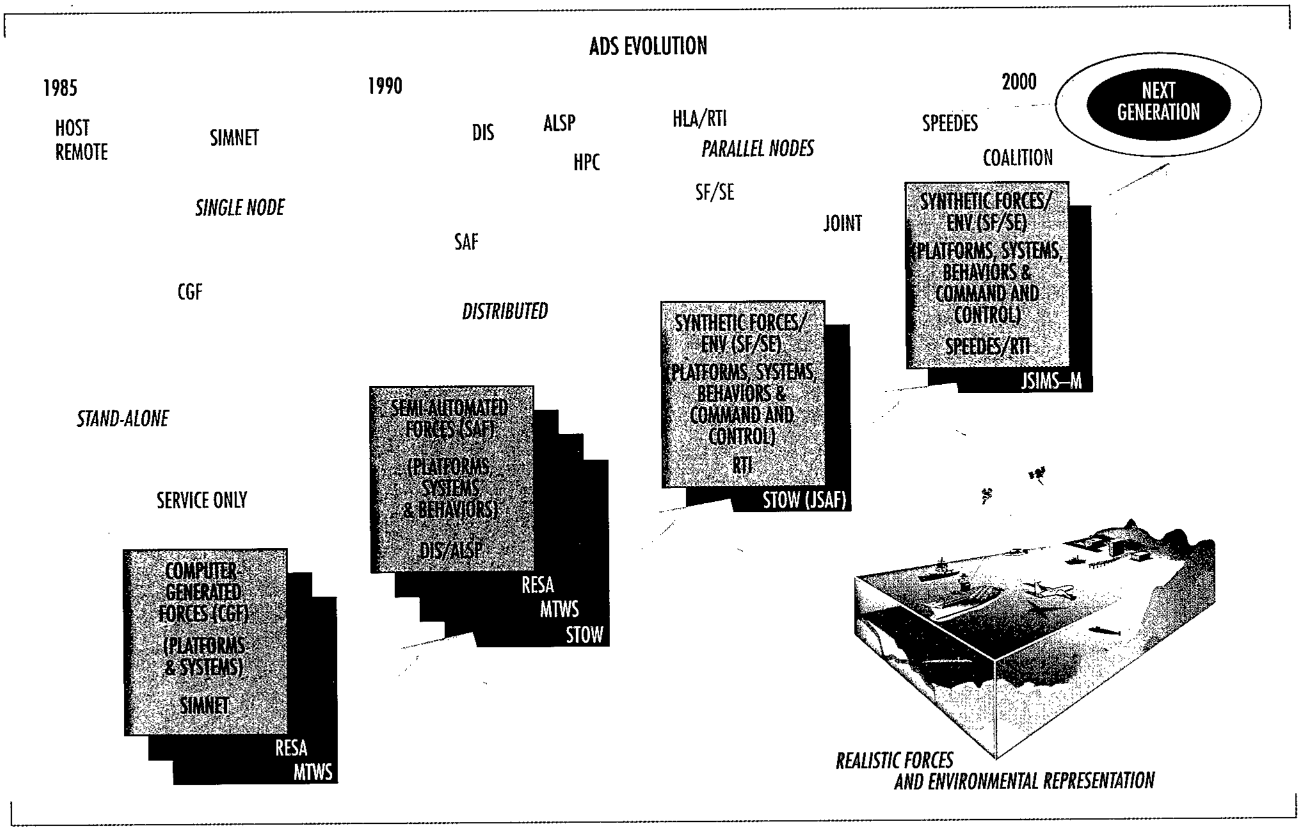

ADS EVOLUTION
1985
HOST
REMOTE
SIMNET
SINGLE NODE
CGF
STAND-ALONE
SERVICE ONLY
COMPUTER GENERATED FORCES (CGF)
(PLATFORMS & SYSTEMS)
SIMNET
RESA
MTWS
1990
DIS
ALSP
HPC
SAF
DISTRIBUTED
SEMI-AUTOMATED FORCES (SAF)
(PLATFORMS, SYSTEMS & BEHAVIORS)
DIS/ALSP
RESA
MTWS
STOW
HLA/RTI
PARALLEL NODES
SF/SE
JOINT
SYNTHETIC FORCES/ ENV (SF/SE)
(PLATFORMS, SYSTEMS, BEHAVIORS & COMMAND AND CONTROL)
RTI
STOW (JSAF)
2000
SPEEDES
COALITION
NEXT GENERATION
SYNTHETIC FORCES/ ENV (SF/SE)
(PLATFORMS, SYSTEMS, BEHAVIORS & COMMAND AND CONTROL)
SPEEDES/RTI
JSIMS–M
REALISTIC FORCES
AND ENVIRONMENTAL REPRESENTATION

FIGURE 4. Advanced distributed simulation evolution through the 1990s and into the 21st century. Leading the way are advancements in network technologies and protocols, computer technologies, modeling representations of forces and environments, and the requirements of a more complex, diverse user community.

Modeling and simulation exposed and evolved these challenges in the 1990s. However, these are just a few of the challenges facing advanced distributed simulation in the 21st century. Next-generation and generation-after-next simulations need to address these questions, and SSC San Diego, with its simulation arsenal, will continue in the forefront of this investigation. (See Figure 4.)

AUTHORS

Elaine C. Allen
BS in Applied Mathematics w/Scientific Programming,
University of California at San Diego, 1990
Current Research: Modeling and simulation development and application.

Kevin P. Adams
BS in Electrical Engineering, Marquette University, 1983
Current Research: Modeling and simulation development.

Charles B. Peters
BS in Marine Biology, San Diego State University, 1974
Current Research: Modeling and simulation application.

Larry J. Peterson
Ph.D. in Computer Science, University of Illinois, 1975
Current Research: Modeling and simulation; high performance computing.

Michael A. Cannon
BA in Economics, University of Rochester, 1970
Current Research: Modeling and simulation development.

Jeffrey S. Steinman
Ph.D. in High Energy Physics, University of California at Los Angeles, 1988
Current Research: Modeling and simulation; high performance computing.

Bruce W. Walter
MS in Systems Management, University of Southern California, 1982
Current Research: Modeling and simulation development.

❖

Douglas R. Hardy
MS in Applied Math/Physics,
Arizona State University, 1985
Current Research: Modeling and simulation development.

"Task-Managed" Watchstanding: Providing Decision Support for Multi-Task Naval Operations

Glenn A. Osga, Karl F. Van Orden, David Kellmeyer, and Nancy L. Campbell
SSC San Diego

ABSTRACT

Watchstanding in shipboard command centers requires U.S. Navy crews to complete time-critical and externally paced task assignments in an accurate and timely manner. Requirements for optimized crew sizes in future ships are driving system designers toward human–computer interface designs that mitigate task and workload demands in a multi-task work environment. The multi-task mission is characterized by multiple concurrent task demands and parallel task goals of varying time duration. Design concepts for a multi-modal watchstation work environment were created that support a variety of crew cognitive and visual requirements during these high-demand missions. Key user support tools include a concept of embedded "task management" within the watchstation software. Early tests of "task-managed watchstanding" have yielded promising results with regard to performance, situation awareness, and work-load reduction. Design concepts are now being transitioned into newer naval systems under SSC San Diego guidance and direction.

INTRODUCTION

Crew size and function allocation in future ships have been recognized as a significant cost factor and therefore have become a performance capability objective for a new class of ships planned for later in this decade [1]. Human performance, driven by a complex, multi-task littoral mission job environment, is the rate-limiting factor for crew optimization. Total task workload must be distributed among a trained crew and controlled in a manner that allows successful performance with minimum risk of mission failure or compromise. Current design practice calls for systematic assignment of tasks (workload) to crew members in a fairly rigid manner—creating periods of high workload or overload for some crew members while others may sit nearly idle with low workload. Crew-size optimization calls for much higher precision in task assignments and workload optimization, with minimum waste in workload capacity as tasks are assigned to the smaller crew.

In 1996, the Multi-Modal Watchstation (MMWS) project was initiated to investigate design concepts that would support crew optimization in command centers. An ergonomic, task-centered watchstation was developed (see Figure 1). The design approach first identified user requirements related to the total work environment and task workload drivers. For purposes of this design discussion, we define a "task" as a job activity with the following attributes:

1. A goal-oriented work activity that results in a defined product.
2. Varying in time from seconds to hours, or the entire watch period (6 hours or more).
3. Supportable by computer-based aids (i.e., not physical work or maintenance activities, although such tasks could benefit by using the principles of this design).
4. Supportable by various levels of automation, which are, in some cases, user-selectable and, in others, may be fixed. Thus, levels of task supervision and user/system task sharing are dynamic.
5. May vary from structured, rigid protocols to open-ended, user-defined sequences. Following Rasmussen's hierarchy [2], tasks may include skill-, rule-, or knowledge-based behaviors.

An important aspect of the task-centric approach is the focus on the "total" work environment, which is defined as mission + computer interface + work management tasks. Naval system designers typically focus on the narrow "mission-specific" requirements to derive the specifications of software functional design. They neglect workload derived from human–computer interface task activities such as computer interface control (e.g., graphical user-interface manipulations). Also neglected is the considerable cognitive workload for work planning, task selection, and time or resource management. The human operator must constantly strategize and allocate attention resources across multiple concurrent events. Current designs offer little or no user assistance to reduce this type of workload or to foster efficiency. The MMWS design focus on task management issues led to a definition of estimated task characteristics for a future naval system, such as listed in Table 1 [4]. (See [3] for discussion of the Task Characteristics approach.) These characteristics provided a starting point for watchstation design concepts based on these requirements. Since task requirements were only available at an abstract level for the future ship [5] and no concept of operations existed at this early design phase, several important assumptions were made about the future task environment such as (1) what degree of automation would be available; (2) multitasking would be required for crew optimization across multiple threats and multiple warfare areas: land attack, air defense, and area air defense; (3) cross-training across multiple tasks would be possible; and (4) system design would permit assignment of any task to any crew member at a watchstation, limited only by authority and planned operating procedures. These task and design requirements were then used as a basis to generate preliminary design concepts.

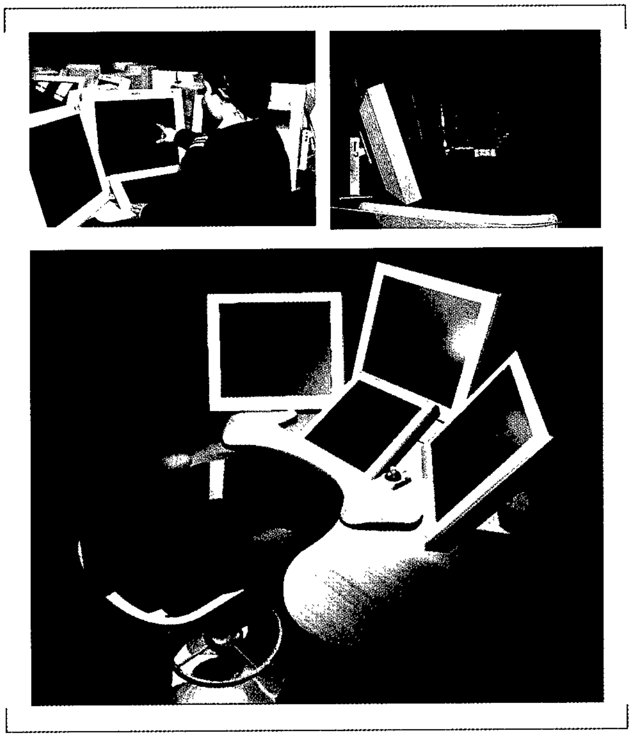


FIGURE 1. Ergonomic Multi-Modal Watchstation Pedestal. The MMWS console was designed to accommodate the 2.5% female through 97.5% male reach envelopes.

PRELIMINARY DESIGN

Each of the concept design requirements was matched with a variety of user-interface aids to support each task type. The design process employed was similar to that noted by Neerincx [6] in which tasks were defined according to their impact on cognitive performance. Specifically, tasks were good candidates for automation support that were judged to be skill- or rule-based. The allocation of task responsibility was considered to be dynamic and user controllable for most tasks. Certain mission tasks better fit the procedural aspects of skill-based behavior (e.g., when the air threat assessment process is completed and the procedural mechanics of issuing warnings or countermeasures become a primary task goal). Design concepts were created to address these projected requirements (Table 1), and examples are listed in Table 2.

The design concept of an "Information Set" was created to contain the "default" or typical information needed to support a task operator. The goal of the design approach was to automate much of the information-seeking task steps. An effective information set would filter pertinent information for the specific task from the visual "noise" or unimportant data. For example, a particular land-attack task in a given geographic sector would require the information set to filter the tactical display to show relevant threats and friendly forces icons. Information sets were defined to contain various graphical user-interface windows such as (1) tactical summary (situation awareness), communications (who to talk or listen to relevant to the task); (2) time and work management (task summary as shown in Figure 2); and (3) amplifying information specific to the task type (e.g., identification [ID] basis information for assessment when issuing a warning). Simple graphic-design rules were developed such as color-filled tactical symbol objects to represent tracks with a pending task and color-outlined symbols to represent no current work in progress.

To address requirements related to depiction of task progress, information formats related to task management were designed. Early concepts addressing air defense task progress were created in 1989 and reported in Osga [7]. Design concepts for the Response Planner Display from the Tactical Decision-Making Under Stress (TADMUS) project were also

TABLE 1. Key task characteristics related to task management requirements.

Task Characteristics Tasks:	**Design Requirement** System should:
May have definable start/stop schedules	Monitor concurrent loading and make schedules visible to user.
Have definable goals	Monitor progress toward goals—offer assistance if needed—report progress toward goals—allow user to modify or create new goals.
Are grouped as parts of overall job role	Provide visual indication of task assignments and task "health."
May be user and/or system invoked	Indicate who has task responsibility. Invoke and "offer" tasks when possible.
Have information and control requirements	Minimize workload to access information or controls.
Are mission- or computer-control focused	Provide full top-down task flow and status for mission tasks with consistent, short multi-modal procedures.
May involve varying levels of automation from full manual to partial to fully automated	Provide visual indication of automation state with supervisory indicators.
May require one or many databases	Do not require the user to know which database for any task. Direct queries automatically.
May require one or many software applications	Require user to know the tasks, not multiple applications—integrate information across the job vs. application.
Will require attention shift between multiple tasks in foreground and background (parallel)	Provide attention management and minimize workload to shift task focus.
Have definable cognitive, visual, and motor workload components	Use task estimates for workload distribution and monitoring among crew members.
Will likely be interrupted	Provide assistance to re-orient progress and resources to minimize working memory load.
Should be consistent from training to field	Provide consistent terms, content, and goals throughout.
Will evolve as missions, systems evolve over the life cycle of the ship	Support reconfiguration of task groupings and addition of new tasks as systems are upgraded.
May be individual or collaborative	Support close proximity and distant collaboration via visual and auditory tools.

TABLE 2. Key MMWS design concepts related to design requirements.

MMWS Design Concepts	**Design Requirement**—System should:
Response Planner/Manager—individual threat response summary. Task Manager Display—composite workload and tasks.	Monitor concurrent loading and make schedules visible to user.
Response Planner/Manager—range-based, single threat summary. Task Manager Display—task summary display.	Monitor progress toward goals—offer assistance if needed—report progress toward goals—allow user to modify or create new goals.
Task Manager Display—team overview and workload indicators.	Provide visual indication of task assignments and task "health."
Task Manager Display—task assignment summary. MMWS context and event monitoring to support task initiation.	Indicate who has task responsibility. Invoke and "offer" tasks when possible.
Multiple display surfaces—maximize visual workspace (within 5 to 95% reach envelope for touch).	Minimize workload to access information or controls.
Task manager task filters. Response Planner procedural list.	Provide full top-down task flow and status for mission tasks with consistent, short multi-modal procedures.
Visual coding of automtion state.	Provide visual indication of automation state with supervisory indicators.
Information sets automatically created.	Do not require the user to know which database for any task. Direct queries automatically.
"Information Sets" assigned to each task.	Require user to know the tasks, not multiple applications—integrate information across the job vs. application.
Multiple displays, task locator icons, intelligent task sorting and priority visual cues.	Provide attention management and minimize workload to shift between task focus.
Visual indication of team workload.	Use task estimates for workload distribution and monitoring among crew members.
Highlight changed information when task is "dormant." Reminders and notes tied to tasks.	Provide assistance to re-orient progress and resources to minimize working memory load.
Top-down task description carried through in display design as well as training curriculum.	Provide consistent terms, content, and goals throughout.
Design TBD	Support reconfiguration of task groupings and addition of new tasks as systems are upgraded.
3-D auditory support to spatialize multiple voice circuits, audio icons and visual/auditory linking of events (audio spatialized to match visual location.	Support close proximity and distant collaboration via visual and auditory tools.

reviewed [8 and 9]. The Response Planner display was used to depict planned response actions in air defense warfare showing task duration and deadlines related to individual air threats. For MMWS, an additional response manager was added for electronic warfare tasks related to uncorrelated electronic-signature reports. Figure 3 (lower part) shows the MMWS "Response Planner/ Manager (RPM)" display concept. This decision support window depicts the major steps in the detect-to-engage

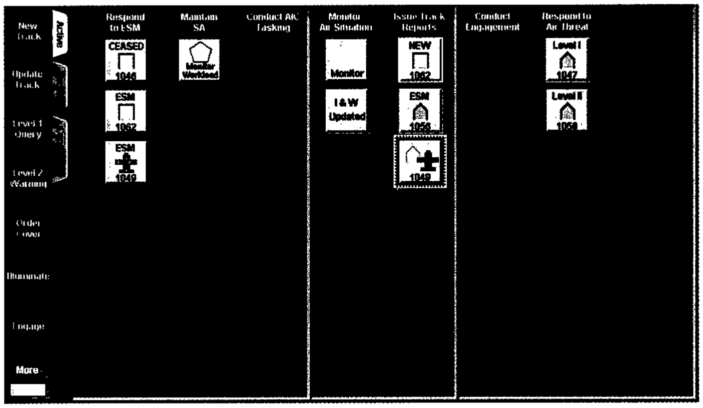


FIGURE 2. MMWS task management display with icons representing tasks awaiting user attention.

sequence that are possible and the ranges at which they might be completed and be in accordance with current response doctrine. Currently recommended task bars are filled white with an unfilled status circle. Previously completed tasks are represented by task bars that are filled black with a green status circle. Tasks that possibly could be triggered if the track maintains its current ID are gray with white letters. Tasks that will not be triggered if the track maintains its current ID are filled in with gray and with gray letters. The task bars are selectable, and the operator can launch a task manually by clicking on them. The RPM window is paired with the Track Profile Window, shown as the upper window in Figure 3. Both windows share a common range-scale from ownship. The track profile window provides a graphical representation of the hooked track's altitude and speed as a function of range from ownship. The altitude trail is color-coded to display the ID history of the track. The speed trail is shown in white. Commercial air transport (COMAIR) ranges are shown colored in purple along both the altitude and speed axis of the graph. Black boxes with white letters displayed along the altitude trail show the tasks performed for that track.

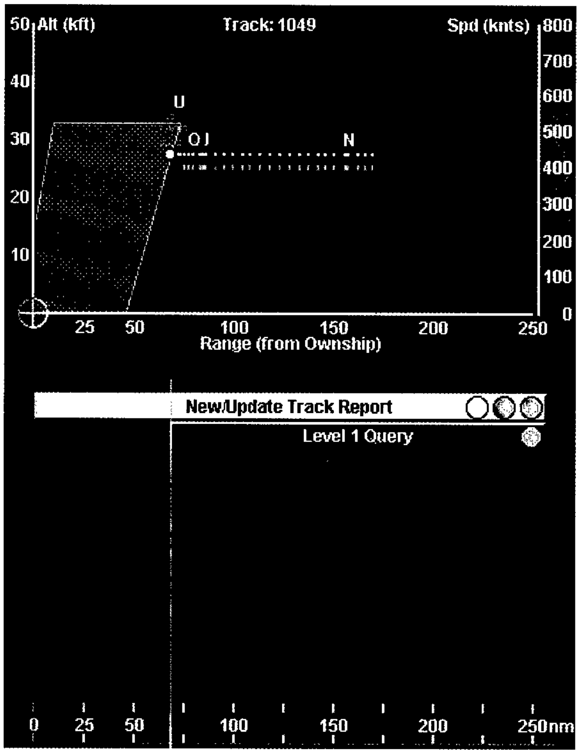


FIGURE 3. Track Profile (upper window) and Response Planner (lower window) displays. This example shows that a New Track report, two Update reports, and a Level 1 Query were previously completed. The track is progressing at a steady altitude (25 kft) and speed (450 knts). The tactical graphics show the weapons envelopes of ownship in teal, and, if applicable, unknown or suspect track possible weapon envelopes are shown in red.

For air defense warfare, the following codes are used on the track profile to display which task was performed:

N = New Track Report issued
U = Update Track Report issued
Q = Level I Query issued
W = Level II Warning issued
V = Visual Identification (VID) ordered
C = Cover ordered
I = Illuminated
E = Engaged

Attention Management

The MMWS design considers the requirement to guide user attention through all phases of the task life cycle. These phases are (1) initiation, (2) orientation, (3) decision, (4) execution, (5) confirmation, and (6) transition. User attention must be directed across and within task activities. Figure 4 illustrates the benefits of consistent color-coding across windows, within a task type. Color-coding for ID illustrates

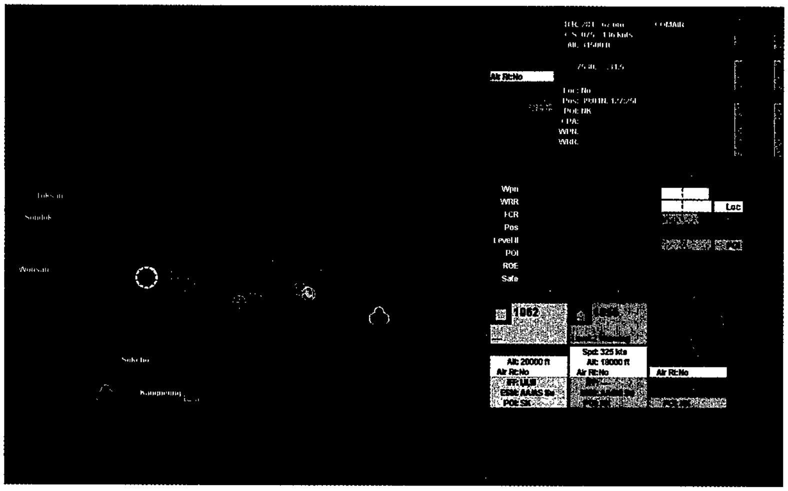


FIGURE 4. Consistent color-coding for ID and improved tactical graphics help to guide user attention and speed visual search tasks. Consistent color-coding across displays aids in information scanning and interpretation. The Track Profile (Figure 3), Amplifying Info, Basis of Assessment, Mini-Amp-Info, and Tactical Displays shown in this figure illustrate the common coding used throughout all windows.

evidence both for and against a given ID assessment. Uniform color represents higher ID certainty while a "rainbow" of color represents less certainty. At a glance, the user can see in each display if there is consistent or conflicting ID evidence, and can quickly assess where the conflicts exist. The Basis of Assessment display provides a history of the changes in ID basis; thus, the user can tell if the data elements are consistent over time or changing. This coding supports efficient visual scanning and task dwell-time optimization. Experts dwell on problem areas such as a "suspect" track with an inconsistent ID basis, and spend less time visually sampling tasks or tracks with consistent information.

Another requirement exists to guide user attention in an efficient manner through multiple tasks. Task detection may be unreliable when the system relies on human vigilance during multi-tasking, and often users are reluctant to drop a non-critical task when a higher priority task appears. There can be a reluctance to leave work unfinished. The MMWS task management system monitors for task-event triggers in the environment. Relative to today's systems, user workload to monitor and trigger tasks should be significantly reduced, allowing attention resources to be allocated for task execution, not task detection. Also, tasks may be categorized with respect to both time and mission urgency. Task management displays have been found to improve judgments about the effect of delays for subtasks and global tasks when problems were introduced into task progress [10]. Results indicated significant performance gains for task management assistance in selecting appropriate response strategies for mission- and time-critical tasks. Automation to support task prioritization of the highest level task improved user efficiency.

Recent usability testing results for the MMWS [11] indicate that visual depiction of time and display scrolling on the task manager were not beneficial during high workload periods. This result led to a revision of the MMWS design concept to allow more tasks to be depicted without scrolling, using visual separation of completed, current, and pending tasks.

Design Testing and Analysis

A critical part of the design and engineering process involves usability testing with fleet participants. Testing involves user hands-on interaction with design items to obtain measures and observations of user training and acceptance, and to identify design items that invoke confusion, error, or slow performance. The goal is to test a few subjects to identify repetitive or common problems across all participants. Significant usability testing has been used to mature the designs in this capability to their current status. Over 75 military and civilian participants were tested from 1997 to 2000 as part of the MMWS development program. Metrics vary in usability testing depending on the focus for the test. During MMWS development, versions 1.x through 5.x were subjected to quantitative measurement. Figure 5 shows the successive changes in question accuracy as scored by accuracy points over four Version 3.x design iterations. Such measures provide an indication of design improvement. Design

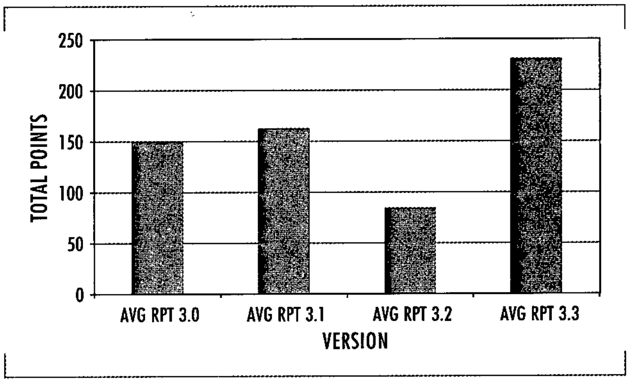


FIGURE 5. Points scored in testing over multiple design versions of MMWS.

comments and workload ratings provide indications of user preference and workload induced by the design and task scenario.

Team performance measurement is a critical design success criterion resulting in quantitative measures of the improvement of the MMWS capability in comparison to existing air defense decision support tools. A realistic air defense problem scenario was used for team performance assessment. The use of the scenario allowed specific comparison of teams using the MMWS Decision Support System (DSS) capability with Aegis teams tested with various Aegis software configurations. This allows a direct assessment of the MMWS DSS capability improvement vs. today's systems. The test was also designed to demonstrate a 50% crew reduction using eight operators in the Aegis team vs. four in the MMWS team. A test goal was to determine if workload and performance could be sustained with reduced crew sizes, such as those proposed for future ship teams. The scenario design was coordinated with Aegis Training and Readiness Command in Dahlgren, VA; subject experts at BCI, Dahlgren, VA; and scientists and engineers at SSC San Diego and Naval Air Warfare Center Training Systems Division (NAWCTSD), Orlando, FL. The scenario was engineered and set in a restrictive warfare environment to foster cognitive workload and decision-making under ambiguous circumstances. Fleet comments at the conclusion of test sessions indicated the scenario was as realistic as other operational test scenarios used in fleet training.

The test scenario contained low and high workload periods and a "coast period" was used in the middle portion of the scenario to allow for further data collection. In the second period, there were more tracks, increased ambiguity of information, and a higher threat situation. The operational parameters for the scenario were defined including:

1. Political Summary
2. Ownship Mission and Tasking
3. Air Tasking Order (ATO) and Carrier (CV) Flight Plan
4. Rules of Engagement (ROE) and Warning/Weapon Status
5. Operational Tasks (OPTASK) Link-ID
6. South Korean Military Tactical Air (TACAIR)
7. OPTASK Air Warfare Plan
8. Call-Signs
9. Operations Order (OPORDER), Warfighting Doctrine and Policy Guidance
10. Communications Assumptions and Plans
11. Location of Air Routes, Return-to-Force Routes, Air Fields and Stations

The scenario was conducted in Condition III steaming, with restrictive ROE and weapons posture for the battlegroup ranging from white/safe to red/tight. Measures included in this study were speed, timeliness, and accuracy (errors of omission or commission). As shown in Figure 6, multiple types of data were collected, including the following:

Timeliness and Accuracy. Collected by viewing video and audiotapes of team actions. Task times were also logged for the enhanced capability version of MMWS.

Efficiency and Workload Capacity. Workload ratings obtained by online scales. Proportion of low criticality tracks addressed by both teams.

Expert Opinion. Subject experts in a review team were assigned to an individual operator. They recorded subject responses to critical track events (25 identified) using the Shipboard Mobile Aid for Training and Evaluation (SHIPMATE) hand-held device.

Situation Awareness. Three probes were conducted during the low and high workload periods. A post-events questionnaire was used during the middle and final coast periods. Questions asked included the following: (1) What are your current tracks of interest? (2) What is your assessment of the intent of Track X? (3) What is your intent with respect to Track Y?

A post-events questionnaire addressed the top tracks of interest and an explanation of the interest. Performance-based inferences also were derived based on tactical response to events in the scenario. Subject-matter experts rated planning, prediction, and critical thinking. The same measures and probes were used for previous Aegis tests [12] and will allow for comparison and measurement of success in this project.

Test Result Highlights

Table 3 shows results indicative of the situation awareness improvement in teams tested using MMWS vs. Aegis crews using legacy equipment. The critical scenario event included a track that appears to be a COMAIR initially, but demonstrates several important kinematic (course, altitude, speed) and other ESM information changes that would warrant increased suspicion. Note in Table 3 that fewer Aegis crews queried or warned the track prior to it attacking the battlegroup, while all MMWS crews did so. The MMWS teams exhibited confidence and awareness in their response actions. With apparently less situation awareness and decision support, Aegis crews used last-second response methods when the air threat launched missiles, while MMWS crews were fully prepared and forewarned. Figure 7 shows that even with a reduced crew size of 50% for the MMWS teams vs. Aegis, the MMWS estimated workload was lower throughout the entire scenario periods tested. Thus, the benefits of the MMWS design included increased situation awareness and performance, with less workload induced on the operating team: a clear win-win situation with respect to performance and workload, therefore reducing mission performance risk.

FIGURE 6. MMWS designs were subjected to individual and team testing in realistic tactical operations.

TABLE 3. Responses of Aegis and MMWS to kinematic changes and ESM events with a critical scenario threat.

	Kinematics	Query/Warning	Engage ASM
Aegis Teams	1 of 8	2 of 8	7 of 8
MMWS V1	6 of 6	6 of 6	6 of 6
MMWS V2	2 of 2	2 of 2	2 of 2

CONCLUSIONS

The MMWS project investigated the design concept of explicitly creating and embedding mission tasks and their associated goals within the visual user interface, using visual priority cues and task progress summaries. The user was assisted throughout the entire task life cycle. Draft task products were prepared for user review, in contrast to the manual workload in visual search, discovery, and task product creation in today's systems. Test results for usability and team performance indicate that the design concepts in MMWS could be a key enabler for crew performance, enabling improved situation awareness and workload reduction. This may be particularly true in multi-tasking missions where workload is externally paced and attention must be distributed across multiple simultaneous tactical events. Task management appears to support work in command and control environments that involve a mixture of rule-, skill-, and knowledge-based tasks. Task management greatly facilitates real-time workload assessment, useful for adaptive automation and re-allocation of functions between team members [13]. Further team-performance research is needed in these complex naval task environments to determine best methods for task distribution and automation monitoring by humans working cooperatively with intelligent task management aids.

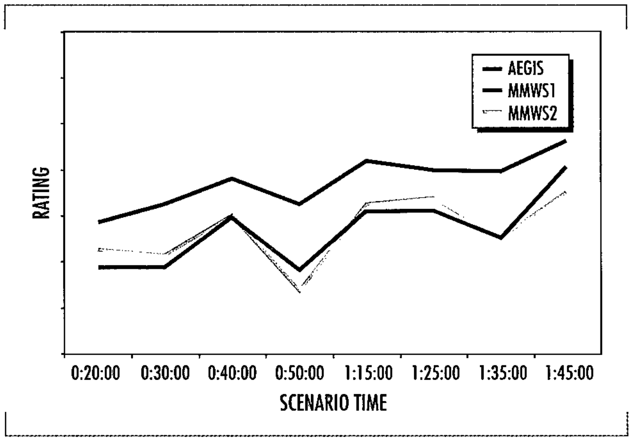


FIGURE 7. Workload levels across scenario periods for Aegis and MMWS as determined by subject-matter-expert ratings.

ACKNOWLEDGMENTS

This paper represents work supported by the Science & Technology Manning Affordability Initiative Program, Office of Naval Research. The content is the authors' and does not reflect official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government.

The authors acknowledge the contribution of software developers and analysts from SSC San Diego and Pacific Science & Engineering Group: Chris Griffith, Alvin Yue, Mike Carlin, Ming Tsai, Neil Acantilado, Dan Lulue, Marty Linville, Jack Houghton, Bryan Croft, Jonathan Richardson, Jimmy Lam, and Robert Yowell.

AUTHORS

CDR Karl F. Van Orden, USN

Ph.D. in Visual Perception and Psychophysics, Syracuse University, 1988

Current Research: Developing and improving visual displays and information management systems to enhance operator performance; developing real-time methods to monitor workload for the Multimodal Watchstation program.

David Kellmeyer

MS in Industrial Engineering, Ohio University, 1992

Current Research: Decision-support systems; supervisory control displays.

Nancy L. Campbell

MS in Electrical Engineering, San Diego State University, 1985

Current Research: Human–system integration; human–computer interface; task-centered design.

Glenn A. Osga

Ph.D. in Human Factors Psychology, University of South Dakota, 1980

Current Research: Human–computer interaction.

REFERENCES

1. Naval Sea Systems Command. 1997. "Operational Requirements Document (ORD) for Land Attack Destroyer DD 21," Document 479-86-97 (Unclassified version), Washington, DC.
2. Rasmussen, J. 1986. *Information Processing and Human–Machine Interaction: An Approach to Cognitive Engineering*, Elsevier, Amsterdam.
3. Meister, D. 1985. *Behavioral Foundations of System Development* (2nd Edition), Robert E. Drieger Publishing Co., Malabar, FL.
4. Osga, G. 1997. "Task-Centered Design," briefing at the Second Multimodal Watchstation Architecture Working Group, (February), San Diego, CA. (Contact author for more information.)
5. Naval Sea Systems Command. 1996. "SC-21 Concept of Operations (CONOPs) DD 21 Ship Requirements," Draft Rev (3), 17 December, Washington, DC.
6. Neerincx, M.A. 1999. "Optimising Cognitive Task Load in Naval Ship Control Centres," *Proceedings of the Twelfth Ship Control Systems Symposium,* October, The Hague, pp. 9–21.
7. Osga, G. 1995. "Combat Information Center Human–Computer Interface Design Studies," TD 2822, Naval Command, Control and Ocean Surveillance Center, RDT&E Division, San Diego, CA.
8. Kelly, R. T., J. G. Morrison, and S. G. Hutchins. 1996. "Impact of Naturalistic Decision Support on Tactical Situation Awareness," *Proceedings of the 40th Human Factors and Ergonomics Society Annual Meeting*, 22 to 26 September, Philadelphia, PA.
9. Morrison, J. G., R. T. Kelly, R. A. Moore, and S. G. Hutchins. 1997. "Tactical Decision Making Under Stress (TADMUS): Decision Support System," IRIS National Symposium on Sensor and Data Fusion, MIT Lincoln Lab, 14 to 17 April, Lexington, MA.
10. St. John, M. and G. Osga. 1999. "Supervision of Concurrent Tasks Using a Dynamic Task Status Display," *Proceedings of the 43rd Human Factors and Ergonomics Society Annual Meeting*, October, pp. 168–172.
11. Kellmeyer, D. and G. Osga, G. 2000. "Usability Testing and Analysis of Advanced Multimodal Watchstation Functions," *Proceedings of the 44th Human Factors and Ergonomics Society Annual Meeting*, 31 July to 4 August, San Diego, CA.
12. Freeman, J., G. Campbell, and G. Hildebrand,. 2000. "Measuring the Impact of Advanced Technologies and Reorganization on Human Performance in a Combat Information Center," *Proceedings of the 44th Human Factors and Ergonomics Society Annual Meeting*, 31 July to 4 August, San Diego, CA.
13. Van Orden, K. F. 2001. "Real-Time Workload Assessment and Management Strategies for Command and Control Watchstations: Current Findings," in Osga, G., K. Van Orden, N. Campbell, D. Kellmeyer, and D. Lulue. 2001. "Design and Evaluation of Warfighter Task Support Methods in a Multi-Modal Watchstation," Technical Document, SSC San Diego, San Diego, CA, in preparation.

❖

Perspective View Displays and User Performance

Michael B. Cowen
SSC San Diego

ABSTRACT

Consoles that use three-dimensional (3-D) perspective views on flat screens to display data seem to provide a natural, increasingly affordable solution for situational awareness tasks. However, the empirical evidence supporting the use of 3-D displays is decidedly mixed. Across an array of tasks, a number of studies have found benefits for 3-D perspective over two-dimensional (2-D) views, while other studies have found rough parity, and still other studies have found 2-D superior to 3-D. Interestingly, many realistic military tasks have complex demands that require both types of views at different points in time. This paper investigates an interface concept called "orient and operate," which employs the advantages of both 2-D and 3-D displays.

INTRODUCTION

Objects and scenes displayed on a flat screen from a 30- to 60-degree perspective viewing angle can convey three-dimensional (3-D) structure and shape. They are increasingly being used in military and civilian occupations such as air warfare, command and control, air traffic control, piloting, and meteorological forecasting. However, they have not been shown to be effective for all tasks. Comparisons between two-dimensional (2-D) (top-down, side) and 3-D (perspective) displays in the literature on a variety of tasks have found mixed results.[*] Several factors have been proposed to account for the differences (see, e.g., [9, 12, and 19]). In an attempt to identify and evaluate the factors important to the effectiveness of the viewing angle, we developed a series of experimental tasks using simple block stimuli (see Figure 1, left) viewed on a non-stereo display. We found that 3-D views were superior for tasks that required understanding the shapes of the blocks, but that 2-D views were superior for tasks that required judging the precise relative position between the blocks and another object (a ball) in the scene [20]. In these experiments, the 3-D view was from 30 degrees with shading, and the 2-D views were from the top, the front, and the side.

We then extended these findings to more complex and naturalistic terrain stimuli. Participants were shown a 7- by 9-mile piece of terrain in either 2-D or 3-D (see Figure 1, right) and asked to perform tasks that required either shape understanding or judging relative position. We again found that 3-D views were superior for the shape understanding tasks, and 2-D views were superior for relative position judgment tasks [21 and 22]. In these experiments, the 3-D view was from 45 degrees with shading, and the 2-D view was a topographic map with color-coded contour lines.

Interestingly, many realistic military tasks have complex demands that require both types of views at different points in time. For these tasks, we propose an interface concept called "orient and operate," which employs the advantages of both 2-D and perspective view displays. A 3-D view can be used initially to orient or obtain an understanding of the layout of

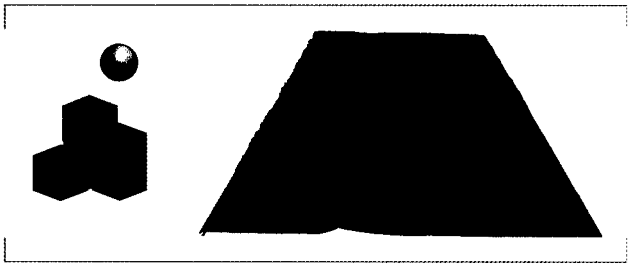


FIGURE 1. Simple block stimuli and terrain stimuli shown in 3-D perspective views.

[*] A number of studies have found benefits for 3-D perspective over 2-D [1, 2, 3, 4, and 5]. Other studies have found rough parity or different results on different measures or tasks [6, 7, 8, 9, 10, 11, and 12] and still other studies have found 2-D superior to 3-D [13, 14, 15, 16, 17, and 18].

background topography and the shape of objects in a scene. Then, a 2-D view can be used to operate on the objects, such as moving them around on the background.

THE GEOMETRY OF 2-D AND 3-D VIEWS

Before continuing, it is useful to understand the basic geometric and functional differences between 2-D views and 3-D views.* One reason 3-D views are good for understanding the general shape of objects and the layout of a scene is that all three spatial dimensions of an object can be seen within a single, integrated view [23]. With a single, integrated view, the user does not need to switch among and integrate information from separate 2-D views to obtain an understanding of the three-dimensional shape of an object or scene. Another reason why 3-D views are good for understanding shape is that natural cues to depth, such as shading, relative size, and texture, can be readily added to an image. Adding these cues can increase the salience of depth in the scene and thereby enhance the sense of a three-dimensional shape. Stereo and motion can also be used to aid the perception of depth,† though these are less commonly used.

One problem for 2-D and 3-D views is that information along the line of sight from the observer into the scene cannot be represented. The reason is that all of the information along a line of sight between the object in the displayed world and the viewer must be represented by the same pixel in a display. In a 2-D top-down or "plan" view, the x and y dimensions are represented faithfully, while the z dimension is lost entirely (see Figure 2). Actually, the x and y dimensions are scaled down in the plan view. "Represented faithfully" means that this scaling is a linear transformation that preserves angles and relative distances in the x-y ground plane so that, for example, parallel lines remain parallel. In the 3-D view, all three spatial dimensions are represented, but the line-of-sight ambiguity remains. Instead of losing one dimension entirely, all three dimensions are foreshortened. The effect of this ambiguity can be seen in Figure 1 (left) where the location of the ball cannot be determined: Is it floating in back of the figure, or is it floating toward the front of the figure?

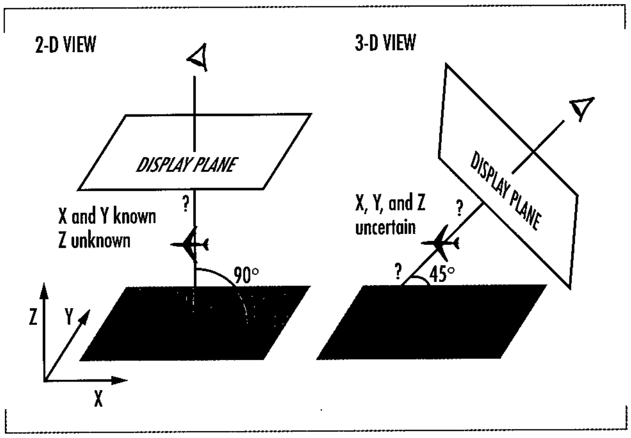


FIGURE 2. Line-of-sight ambiguity makes the location of the aircraft uncertain in different ways, depending on the viewing angle.

A further problem for 3-D views is distortion in the representation of distances and angles. Some distortions result from foreshortening, which increases as the viewing angle drops from directly top-down to ground level. This distortion can cause the sides of a square to appear shortened and the right angles to appear acute or obtuse, as seen in Figure 2. Other distortions result from perspective projection, which causes distances in the x and z dimensions to scale linearly (i.e., a linear perspective), but distances in the y dimension to scale nonlinearly. Due to this distortion, parallel lines appear to converge toward the vanishing point, as can be seen in Figure 1 (right). Perspective projection is, in fact, a cue to depth, but it works by distorting distances and angles. It can make depth more salient in an image, but at the price of making precise measurements more difficult.

* Sedgwick [24] provides a thorough description of 3-D views and perceptions of space.

† See our report [25] for a description of depth cues.

Antenna Placement Experiment

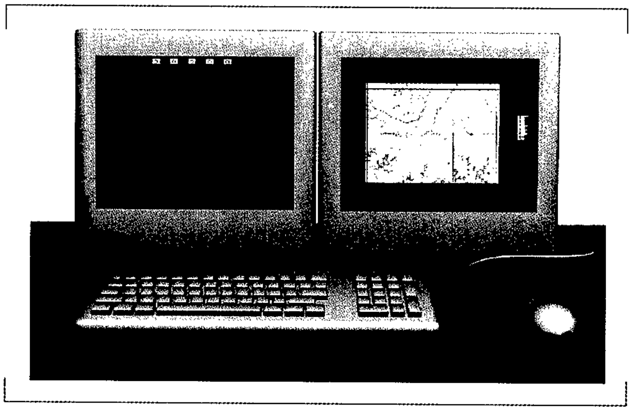


FIGURE 3. Side-by-side condition from antenna experiment: 3-D perspective view map (left) and 2-D top-down topographic view map (right).

Here, we will discuss an experiment that evaluates our interface concept of "orient and operate" using a relatively detailed operational military task. In this experiment, participants were shown a terrain map that contained two fixed antennas (a source and terminal), several enemy unit locations, and a set of antennas to be placed on the map to establish line-of-sight communications. The task was to create a chain of antennas across the map to connect the source and terminal antennas. The antennas had to be within line of sight of each other while remaining concealed from the enemy units. Participants positioned antennas simultaneously out of sight of the enemy, but in line of sight and range of other antennas, thereby creating a chain of antennas across the map. One group of participants viewed only the 2-D topographic map. Another group received only the 3-D view, and a third group received both views, side by side. In the side-by-side condition, the two views were visible to the participant on separate monitors: a 3-D "orient" view and a 2-D "operate" view (see Figure 3). The antennas were constantly visible on both views, even as they moved, so participants could look at either view at their discretion. Participants were timed to complete a series of nine problems.

It was not entirely clear which type of view would prove better for making these precision judgments. In previous work [21], we used line-of-sight judgments as a shape understanding task and found that 3-D views were superior. Participants viewed a terrain segment in either a 2-D top-down topographic view or a 3-D perspective view and judged whether or not there was a line of sight between two points on the terrain. This task appeared to require only a very general gestalt understanding of the terrain—whether a large mountain or range of hills was obstructing the line-of-sight view. In contrast, placing antennas on a map to create an unbroken chain of line-of-sight communications while keeping them out of sight of enemy units may require judgments that are far more precise.

We found that performance with 2-D maps was, in fact, much better than performance with 3-D maps. Our interpretation is that routing of antennas requires placements of units just in or out of lines of sight, and these precise judgments are facilitated by the 2-D view with its faithful representation of space. Interestingly, performance in the side-by-side condition proved to be even better than performance in the 2-D condition. Our interpretation is that some aspects of the antenna task, namely, orientation aspects, were still better performed in 3-D.

We investigated this interpretation in a follow-on experiment. From observations of participants, we found that the 3-D views appeared to be useful at various points throughout the task to help interpret the 2-D topographic views, and that the 3-D views were especially important toward the beginning of the task for determining a basic route. We believe that the ability of the 3-D views to naturally and easily convey shape makes them useful for finding canyons and hills that could be used to build a route through the terrain. This idea fits with our concept of "orient and operate," wherein the user first orients to a scene using a 3-D view and then switches to a 2-D view to perform fine-tuned operations on the scene.

In the follow-on experiment, called "pick-a-path," participants were shown three potential routes across the terrain for constructing their chain of antennas (see Figure 4). One of the three routes was much more promising than the other two, in that it followed canyons and skirted hilltops to remain out of enemy lines of sight. Participants were shown the terrain and routes in either 3-D perspective views or 2-D topographic views. "Pick-a-path" performance was found to be much faster for the 3-D perspective views than for the 2-D views.

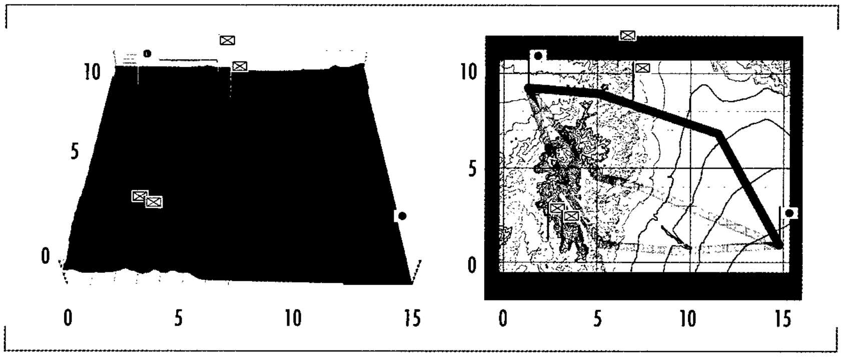


FIGURE 4. An example 3-D map and the equivalent 2-D map from "pick-a-path."

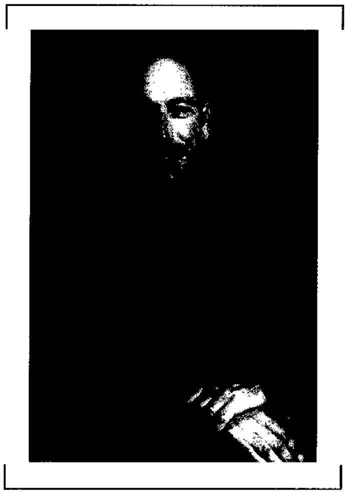
We concluded that the ability to select a path on a terrain map depends not only on the viewing perspective (e.g., 2-D, 3-D), but also on how precise the route needs to be. Initial path planning benefited from a 3-D view while the actual routing of the antennas benefited from a 2-D view. The 3-D view was better able to convey terrain shapes, and the 2-D view was better able to convey where two objects needed to be placed to solve the tactical problem. We recommend using 3-D for initial path planning and 2-D for object placement, supporting our "orient and operate" display design paradigm: Users should orient to a scene using a 3-D perspective view and then operate on the objects in the scene using a 2-D view.

Further supporting "orient and operate," we found that participants performed the best when provided with both 2-D and 3-D views. However, the effect was of small magnitude, and we believe that more improvement is possible. Placing views side by side may not be sufficient for creating an effective suite of displays. Moving from one view to the other requires considerable re-orientation to the scene by the user. Methods are needed for improving the correspondences between objects in the views that alleviate the effects of re-orientation. The concept of visual momentum [26] may offer ideas, such as the use of natural and artificial landmarks, for improving the correspondence. Investigation of these and other concepts is currently underway.

Our antenna placement experiments extended our program of research on how to improve perception of object shape, position, and location to a more complex and applied operational domain. In this domain, we found considerable support for our basic distinction for using 3-D perspective views for shape understanding and for using 2-D views to judge relative position of objects. Applying this framework, we are currently building several "orient and operate" prototypes for use in real-world military display systems.

REFERENCES

1. Bemis, S. V., J. L. Leeds, and E. A. Winer. 1988. "Operator Performance as a Function of Types of Display: Conventional Versus Perspective," *Human Factors*, vol. 30, 163–169.



Michael B. Cowen

Ph.D. in Cognitive Psychology, Claremont Graduate University, 1991

Current Research: Perceptual and cognitive information processing; human–machine interface; advanced training environments.

2. Ellis, S. R., M. W. McGreevy, and R. J. Hitchcock. 1987. "Perspective Traffic Display Format and Airline Pilot Traffic Avoidance," *Human Factors*, vol. 29, pp. 371–382.
3. Hickox, J. G. and C. D. Wickens. 1999. "Effects of Elevation Angle Disparity, Complexity and Feature Type on Relating Out-of-Cockpit Field of View to an Electronic Cartographic Map," *Journal of Experimental Psychology: Applied*, vol. 5, pp. 284–301.
4. Liter, J. C., B. S. Tjan, H. H. Bulthoff, and N. Kohnen. 1997. "Viewpoint Effects in Naming Silhouette and Shaded Images of Familiar Objects," Technical Report No. 54, Max-Planck-Institut fur Biologische Kybernetik, Tubingen, Germany.
5. Naikar, N., M. Skinner, Y. Leung, and B. Pearce. 1998. "Technologies for Enhancing Situation Awareness in Aviation Systems: Perspective Display Research," Technical Report, Air Operations Division, Aeronautical and Maritime Research Laboratory, Defence Science and Technology Organisation (DSTO), Melbourne, Victoria, Australia.
6. Andre, A. D., C. D. Wickens, L. Moorman, and M. M. Boschelli. 1991. "Display Formatting Techniques for Improving Situation Awareness in Aircraft Cockpit," *International Journal of Aviation Psychology*, vo1. 1, pp. 205–218.
7. Baumann, J. M., S. I. Blanksteen, and M. T. Dennehy. 1997. "Recognition of Descending Aircraft in a Perspective Naval Combat Display," *Journal of Virtual Environments*, http://www.hitl.washington.edu/scivw/JOVE/articles/mdjbsb.html
8. Burnett, M. S. and W. Barfield. 1991. "Perspective versus Plan View Air Traffic Control Displays: Survey and Empirical Results," *Proceedings of the Human Factors and Ergonomics Society 35th Annual Meeting*, pp. 87–91.
9. Haskell, I. D. and C. D. Wickens. 1993. "Two- and Three-Dimensional Displays for Aviation: A Theoretical and Empirical Comparison," *International Journal of Aviation Psychology*, vol. 3, pp. 87–109.
10. Van Breda, L. and H. S. Veltman. 1998. "Perspective Information in the Cockpit as a Target Acquisition Aid," *Journal of Experimental Psychology: Applied*, vol. 4, pp. 5–68.
11. Wickens, C. D., C. Liang, T. Prevett, and O. Olmos. 1996. "Electronic Maps for Terminal Area Navigation: Effects of Frame of Reference and Dimensionality," *International Journal of Aviation Psychology*, vol. 6, pp. 241–271.
12. Wickens, C. D. and T. T. Prevett. 1995. "Exploring the Dimensions of Egocentricity in Aircraft Navigation Displays," *Journal of Experimental Psychology: Applied*, vol. 2, pp. 110–135.
13. Boyer, F., M. Campbell, P. May, D. Merwin, and C. D. Wickens. 1995. "Three-Dimensional Displays for Terrain and Weather Awareness in the National Airspace System," *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*, pp. 6–10.
14. Boyer, B. S. and C. D. Wickens. 1994. "3-D Weather Displays for Aircraft Cockpits," Technical Report ARL-94-11/NASA-94-4, Aviation Research Lab, Savoy, IL.
15. O'Brien, J. V. and C. D. Wickens. 1997. "Free Flight Cockpit Displays of Traffic and Weather: Effects of Dimensionality and Data Base Integration," *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting*, pp.18–22.
16. Wickens, C. D., M. Campbell, C. C. Liang, and D. H. Merwin. 1995. "Weather Displays for Air Traffic Control: The Effect of 3-D Perspective," Technical Report ARL-95-1/FAA-95-1, Aviation Research Lab, Savoy, IL.

17. Wickens, C. D. and P. May. 1994. "Terrain Representation of Air Traffic Control: A Comparison of Perspective with Plan View Displays," Technical Report ARL-94-10/FAA-94-2, Aviation Research Lab, Savoy, IL.

18. Wickens, C. D., S. Miller, and M. Tham. 1996. "The Implications of Data-Link for Representing Pilot Request Information on 2-D and 3-D Air Traffic Control Displays," *International Journal of Industrial Ergonomics*, vol. 18, pp. 283–293.

19. Van Orden, K. F. and J. W. Broyles. 2000. "Visuospatial Task Performance as a Function of Two- and Three-Dimensional Display Presentation Techniques," *Displays*, vol. 12, pp. 17–24.

20. St. John, M. and M. B. Cowen. 1999. "Use of Perspective View Displays for Operational Tasks," TR 1795, SSC San Diego, San Diego, CA, http://www.spawar.navy.mil/sti/publications/pubs/tr/1795/tr1795.pdf

21. St. John, M., H. M. Oonk, and M. B. Cowen. 2000. "Using Two-Dimensional and Perspective Views of Terrain," TR 1815, SSC San Diego, San Diego, CA, http://www.spawar.navy.mil/sti/publications/pubs/tr/1815/tr1815.pdf

22. St. John, M., H. S. Smallman, H. M. Oonk, and M. B. Cowen. 2000. "Navigating Two-Dimensional and Perspective Views of Terrain," TR 1827, SSC San Diego, San Diego, CA, http://www.spawar.navy.mil/sti/publications/pubs/tr/1827/tr1827.pdf

23. Wickens, C. D., and C. M. Carswell. 1995. "The Proximity Compatibility Principle: Its Psychological Foundation and Relevance to Display Design," *Human Factors*, vol. 37, no. 3, pp. 473–494.

24. Sedgwick, H. A. 1986. "Space Perception," *Handbook of Perception and Human Performance* (K. R. Boff, L. Kaufman, and J. P. Thomas, eds.), Wiley, New York, NY, vol. 1, pp. 2101–2157.

25. Smallman, H. S., E. Schiller and M. B. Cowen. 2001. "Track Location Enhancements for Perspective View Displays," TR 1847, SSC San Diego, San Diego CA, http://www.spawar.navy.mil/sti/publications/pubs/tr/1847/tr1847.pdf

26. Woods, D. D. 1984. "Visual Momentum: A Concept to Improve the Cognitive Coupling of Person and Computer," *International Journal of Man–Machine Studies*, vol. 21, pp. 229–244.

❖

Decision Support Displays for Military Command Centers

Jeffrey G. Morrison
SSC San Diego

ABSTRACT

This paper summarizes work on interface design requirements for decision support tools and for command centers at the Commander, Joint Task Force (CJTF) level. These tools include a "knowledge wall" for decision-makers and multi-modal workstations for the liaison officers who maintain the summary situation displays for each functional area, enabling a new concept of operations based on enhanced situation awareness throughout the command team.

INTRODUCTION

For over 10 years, SSC San Diego, with sponsorship from the Office of Naval Research (ONR), has been striving to develop improved displays based on decision support technology for military decision-making. At the center of this effort has been the Tactical Decision-Making Under Stress (TADMUS) project and its successors. The TADMUS project was spawned by the 1988 USS *Vincennes* (CG 49) incident, in which an Aegis cruiser, engaged in a littoral peacekeeping mission, shot down an Iranian Airbus after mistaking it to be a tactical threat. Investigations following the incident suggested that stress may have affected decision-making and that the effects of stress were not well understood. The TADMUS project was established to address these concerns and to develop improved decision support tools for use by command decision-makers.

TADMUS developed a series of prototype decision support tools that came to be embodied as the integrated Decision Support System (DSS), (Figure 1). The DSS research showed that when tactical decision-makers had the prototype DSS available, significantly fewer communications were needed to clarify the tactical situation, significantly more critical contacts were identified earlier, and a significantly greater number of defensive actions were taken against imminent threats. Furthermore, false alarms were reduced by 44%, and correct detection of threat tracks

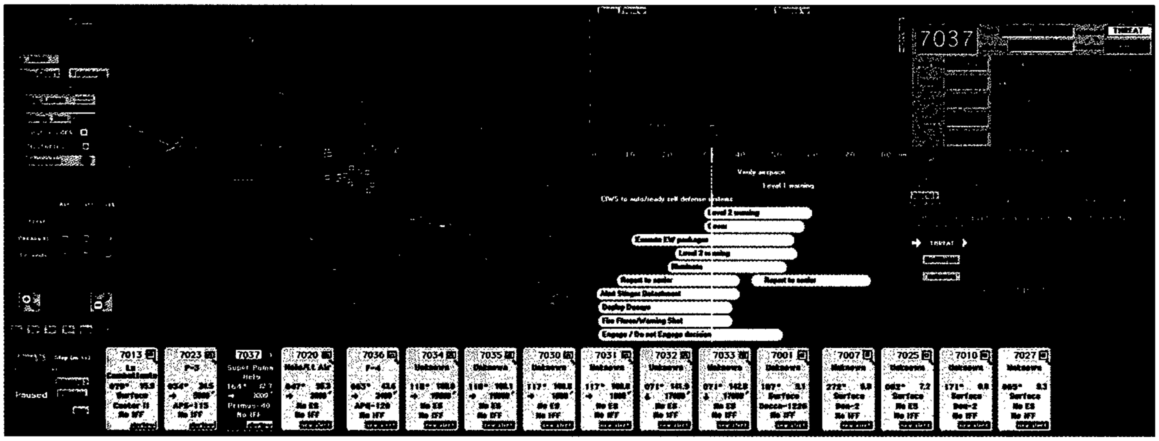


FIGURE 1. The TADMUS Decision Support System.

increased by 22%. These findings suggest that the prototype DSS enhanced the commanders' awareness of the tactical situation, which in turn contributed to greater confidence, lower workload, reduced errors in adherence to rules of engagement, and more effective performance.

The Chief of Naval Operation's Strategic Studies Group XVI report "Command 21—Speed of Command" recognized the significance of the TADMUS work and stated that its results were more broadly applicable. The Group concluded that

- Fleet decision-makers are faced with too much data and not enough information.
- Fleet information systems are often not designed to support the decision-makers.
- Reduced manning requirements and complex mission requirements will further exacerbate the problem.

One of the key recommendations to come out of the Command 21 report was that decision support technology developed in the TADMUS project should be extended from single ship combatants to higher echelons of command. The Command 21—Decision Support for Operational Command Centers (Command 21) project is addressing this recommendation by conducting research into the unique requirements of decision-making within military operational command centers.

The initial Command 21 work with Second and Third Fleet command ships has suggested that (1) collaboration is problematic in these command centers, and (2) commercial off-the-shelf (COTS) collaboration tools often are not as useful as might be expected. Military decision-makers were found to engage in "asynchronous collaboration," where each was working on different parts of a common problem in their own space and their own time, and as a result, each having their own decision cycle. This situation is different from traditional "synchronous" collaboration, such as the "brainstorming" or group problem solving found in the business world. Staff-wide synchronization is largely achieved when briefings are given to the assembled staff at watch-turnover. A central premise for Command 21 is that "Speed of Command" can only be achieved when it is not necessary to stop and brief command decision-makers so that they can be fully informed as a basis for deciding what actions to take. The Command 21 project has developed a concept of operations for sharing information that incorporates unique, Web-enabled collaboration "push" tools to provide all decision-makers ready access to the best available data at all times.

One Command 21 tool is the "knowledge wall," shown in Figure 2. The wall features a series of windows incorporating decision support tools tailored to the Commander Joint Task Force (CJTF), as well as windows with "summary status" information being "pushed" from the anchor desks used by liaison officers (LNOs) representing the various

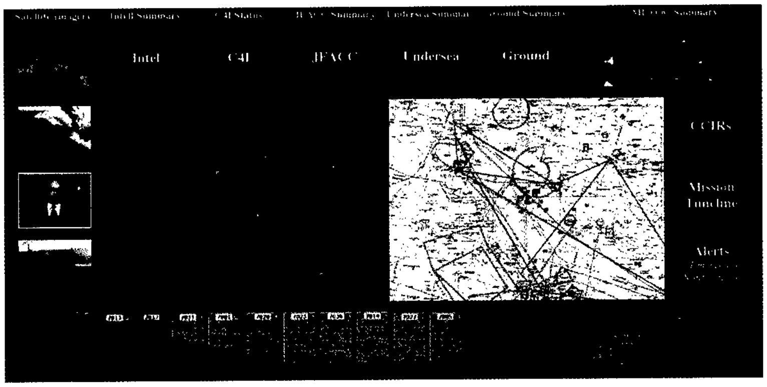


FIGURE 2. Command 21 knowledge-wall vision.

CJTF departments. The battle watch captain in charge of the command center can choose which aspects of the situation to focus on by moving relevant content to the center of the wall and drilling down into deeper levels or related information.

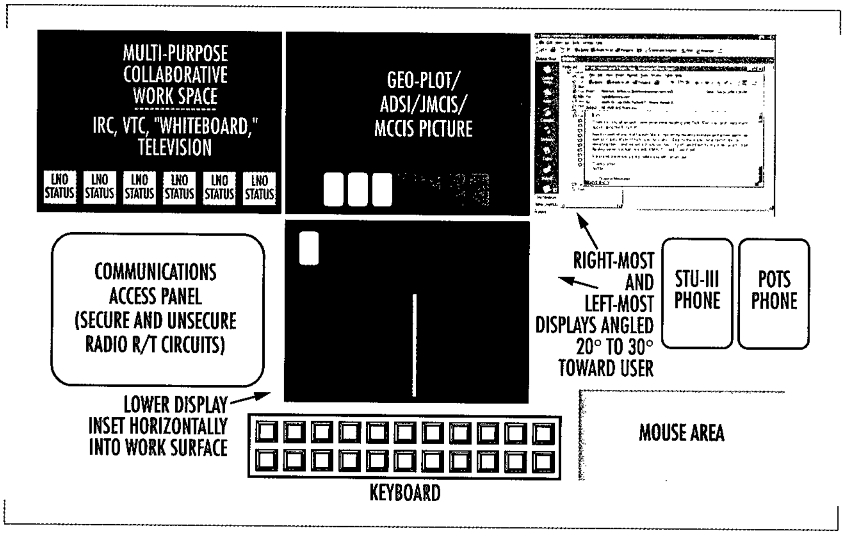

MULTI-PURPOSE COLLABORATIVE WORK SPACE
IRC, VTC, "WHITEBOARD," TELEVISION
LNO STATUS
LNO STATUS
LNO STATUS
LNO STATUS
LNO STATUS
LNO STATUS
GEO-PLOT/ ADSI/JMCIS/ MCCIS PICTURE
COMMUNICATIONS ACCESS PANEL (SECURE AND UNSECURE RADIO R/T CIRCUITS)
RIGHT-MOST AND LEFT-MOST DISPLAYS ANGLED 20° TO 30° TOWARD USER
STU-III PHONE
POTS PHONE
LOWER DISPLAY INSET HORIZONTALLY INTO WORK SURFACE
MOUSE AREA
KEYBOARD

FIGURE 3. Knowledge-desk concept.

A watchstation being developed for the DD 21 (21st Century Destroyer) as part of the ONR Manning Affordability Advanced Technology Demonstration could be adapted as a "knowledge desk" to allow LNO collaboration. The knowledge desk uses software tools (COTS and information-push Web applications) together with computer display hardware to enable the operator to create and publish value-added information to the Web. Figure 3 shows a conceptual version of the knowledge-desk operator console. It consists of an integrated "desktop" spread across four different display surfaces. The top-right display is dedicated to routine office tasks such as preparing briefs, processing e-mail, writing memos, etc. The top-center display is dedicated to providing the tactical situation "big picture" tailored to the user's decision-making needs. The bottom-center display is a dedicated place for monitoring the execution of an operational plan. The top-left display is a tool explicitly designed to facilitate sharing information. The concept uses templates to "push" information from the operator to a Web site viewable by the rest of the command staff. The information "pushed" consists of worksheets, forms, and prompts to others on the command staff that would facilitate their understanding information relevant to their decision-making tasks. The software tools cause the information pushed to be formatted in a manner that others would recognize and understand, and published to a shared database in the Web environment.

The development of the knowledge wall was greatly accelerated through its use as part of the Global 2000 wargame. The objective of this game was to explore how the elimination of "stove pipe" command and control systems (i.e., "network-centric warfare") might change the way we perform military missions. The wall was designed using COTS hardware and software capabilities that exist today so as to minimize development costs, and therefore differs from the original Command 21 knowledge-wall vision. Figure 4 shows the knowledge wall as installed in the Joint Command Center at the Naval War College.

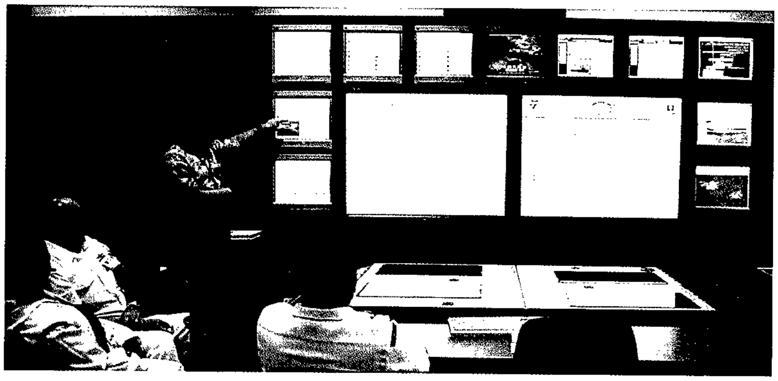
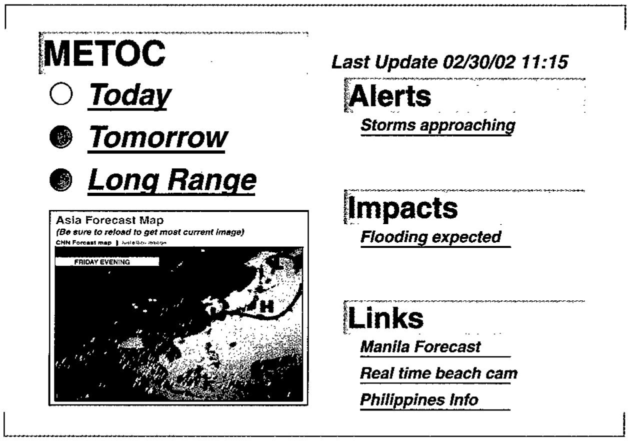


FIGURE 4. Global 2000 wargame knowledge wall.

The knowledge-wall hardware consists of a dual-processor Information Technology for the 21st Century (IT-21)-compliant workstation using three 4-port Appian Jeronimo Pro COTS video boards. The knowledge-wall display is made up of ten 21-inch CRTs and two SmartBoard rear-projection large-screen displays with internal liquid-crystal display (LCD) projectors. The displays operate as a single, integrated digital desktop, where each physical display has a resolution of 1024 by 768 pixels. This creates a digital desktop of 6144 by 1536 pixels. An additional CRT is dedicated to video and video teleconferencing requirements.

The peripheral displays are intended to provide summary information for each of 14 functional areas of the CJTF command identified through knowledge engineering with the staffs of the U.S. Navy Third Fleet, Carrier Group One, and Carrier Group Three. Each summary display is formatted consistently by using a template-authoring tool that facilitates the creation of, and linking to, a variety of Web content without the operator responsible for producing content having to know hypertext mark-up language (HTML). Additional authoring tools were provided to facilitate the creation and publishing of map-based tactical data. All pages are implemented as HTML pages on a common server, with numerous links to more detailed pages for supplemental information.

Figure 5 shows how the information might look in a representative summary display. The title line indicates the functional area described by the display. The "stop lights" in the top-left quadrant are intended to be viewable from 15 to 20 feet away, and indicate the status of activities in various time frames. Light colors indicate the severity of the alerts in terms of their deviation from the plan. The bottom-left quadrant provides space for a summary graphic or multimedia object. The right side of the screen provides space for amplifying links/headlines. The "Alerts" section describes specific problems within this domain/functional area that might be of interest to others. The "Impacts" links describe the impacts of alerts in terms of effects on other functional areas. The "Links" area allows access to reference and supplemental material. Any text or graphic in the page may be linked to a more detailed Web page.



METOC

Last Update 02/30/02 11:15

○ Today

● Tomorrow

● Long Range

Asia Forecast Map
(Be sure to reload to get most current image)
CNN Forcast map

FRIDAY EVENING

Alerts
Storms approaching

Impacts
Flooding expected

Links
Manila Forecast
Real time beach cam
Philippines Info

FIGURE 5. Representative summary display.

The Global 2000 wargame substantially validated the case for the use of Web-enabled decision support and collaboration tools as a means to "Speed of Command" and network-centric warfare. At the start of the game, it was argued that speed of command meant not having to stop to have a situation briefing to figure out what was known across the staff. By using the knowledge wall and a number of information technology collaboration tools, not one staff briefing was required through 8 days of game play. The wall was used extensively, with 30 to 70 unique summary pages being accessed each hour.

Both the TADMUS and Command 21 projects have empirically demonstrated how the application of decision support technology and effective human factors can improve military decision-making by turning data into meaningful information presented where, when, and the way it is needed. The Global 2000 wargame showed how network-centric warfare, in combination with decision support and a Web-enabled command and control architecture can move tomorrow's military to "knowledge-centric warfare."

❖

Jeffrey G. Morrison

Ph.D. in Psychology, Georgia Institute of Technology, 1992

Current Research: Decision support technology; knowledge engineering/management; collaboration.

Development of Wearable Computing, Advanced Visualization, and Distributed Data Tools for Mobile Task Support

Steve Murray
SSC San Diego

ABSTRACT

This paper presents an overview of SSC San Diego projects to develop novel information systems based on wearable computing, advanced visualization, and wireless technologies. User-centered design of these devices, to include proper principles of human perception and cognition, enables individuals to interact seamlessly with information and with other personnel regardless of their physical location. The results are better decision-making and shortened timelines for task completion.

INTRODUCTION

New approaches to on-the-job information support are being made possible by advances in wearable computing, hand-held information devices, and wireless communications technologies. Expanded data-storage capacity, innovative visual displays, and small lightweight packaging provide many choices for the design of systems that enhance information access, decision-making, and communication among sailors or Marines regardless of their location.

While commercial products can be assembled to accommodate a variety of purposes, an enterprise-level perspective is still required to realize their full potential. SSC San Diego has supported this need by integrating diverse commercial technologies, mapping them to user applications, adding design or functional improvements as appropriate, and conducting impartial performance testing of the resulting systems. SSC San Diego's goal is to ensure the smooth integration of commercial products into capable and robust military systems that support new operational capabilities.

ENABLING TECHNOLOGIES

Mobile information tools continue to emerge from industry at an accelerating rate. Critical technologies for enabling mobile support include improved computing resources, innovative information displays, interaction tools optimized for portability, a range of small imaging sensors, and a wireless communications infrastructure. These technologies can provide the user with responsive, easily accessible task and decision support at virtually any work location. The quantity and variety of these new products, however, only highlight the essential engineering tasks of system integration and testing to ensure that technology investments are ultimately realized as practical enhancements to naval capabilities. Although such tools may work well independently, it is their combined interaction that provides major advances in mission effectiveness.

Computing Resources

Computing and storage power for wearable or hand-held computers expands at roughly the same rate as desktop units, owing to the increased interest in these portable devices for an ever-widening range of industrial jobs. Typical commercial systems feature Pentium II CPUs in the 233-MHz class, with random access memory (RAM) resources of 160 MB and integrated hard-drive storage of 8 GB. Many vendors have already announced systems with greater power.

Information Displays

High-resolution color displays (e.g., 640 x 480 pixels), readable in both bright and dim light, are now available in hand-held and head-worn variants. While hand-held systems are most common, head-mounted displays (HMDs) support task performance in unique ways. In particular, HMD information is always available in the field of view so the user does not need to look away from the workspace. Some systems feature "see through" capability, where information is presented on a semitransparent surface or on the lenses of eyeglasses. If additional sensors are added to the system to track head position, displayed information can be synchronized, or registered, with the real-world scene, much like a pilot's head-up display (HUD). This approach is known as "augmented reality," and current applications include labeling and explanations of equipment parts, visualization of subassemblies that cannot be directly seen, animations of component operation, and sequential cueing of procedures as they are performed. SSC San Diego researchers have developed new display metaphors for effectively presenting information on HMDs—with special emphasis on augmented reality—and have conducted systematic user testing to establish the most appropriate allocation of information between hand-held and head-worn displays. In addition, SSC San Diego has generated inexpensive concepts for head tracking required to support practical augmented reality displays.

Interaction Tools

Miniaturized keyboards, keypads, and mouse tools are already familiar to users of portable computers, although stylus tools and speech recognition are becoming more common due to personal digital assistant (PDA) popularity and the growing need for hands-free computer interaction in offices. SSC San Diego has tested each of these technologies and has, additionally, developed gesture control methods (i.e., computer interaction using hand and finger movements with specially instrumented gloves) for interacting with information on HMDs.

Sensors and Imaging Tools

The utility of mobile information devices is clearly enhanced when they are equipped with sensors that capture data about the work environment (to document a task or to share visualization with others) or when they are equipped with sensors that extend human senses in hazardous situations. Video and still cameras are commonly used in industrial settings to support maintenance collaborations with remote technicians, and both military and civilian communities have employed thermal and low-light sensors during firefighting and surveillance tasks. SSC San Diego engineers are exploring the roles of such sensors in a variety of field, ship, and shore settings through user interviews and job analysis.

Information Sharing

Whether recording data on site, transmitting data to another site, or accessing remote data resources, essentially all naval jobs involve information sharing. Portable information tools on the commercial market typically offer some form of sharing through physical transfer (e.g., computer docking), or through radio, infrared, modem, or cell phone connectivity. Most recently, wireless local-area network (LAN) technologies—and Internet-based communications methods—have become a primary focus of fleet interest for distributed information exchange aboard ship. Internet-based communications are useful for linking networks of people and data sources with each other. SSC San Diego is actively involved with ship- and shore-based wireless LAN systems, and has designed innovative extensions to Internet communications protocols that support the unique demands of mobile, intermittent connectivity (such as lost or unreliable communications nodes, retransmission of unacknowledged data, etc.).

SSC SAN DIEGO DEVELOPMENT ROLE

There is no shortage of portable, yet potent information technologies to support mobile Marines and sailors. Operational effectiveness of new systems, however, must be preceded by a development process that starts with examination of user task and information needs, moves through informed selection and integration of component technologies, and concludes with field validation testing. Given that most technologies now originate from the commercial sector, execution of this process represents the essential "value added" contribution of SSC San Diego engineering. Two projects that illustrate this SSC San Diego development role in wearable computing technologies are the Advanced Interface for Tactical Security (AITS) and the Virtual Technical Data System (VRTDS).

Advanced Interface for Tactical Security (AITS)

The AITS project—an initial SSC San Diego effort in mobile computing and visualization—was intended to support field soldiers. Specifically, the Defense Threat Reduction Agency (DTRA) charged SSC San Diego with developing an intuitive information interface for U.S. Army security system operators (although units of all the military services perform similar missions). These personnel monitor sensors of diverse types are placed around the perimeter of a protected area. When sensors detect an intrusion, security operators must quickly orient themselves and interpret the nature of any threat. The AITS design effort began with observations and interviews of several security units, and proceeded based on documented user information needs.

AITS is based on a commercial wearable computer with both HMD and backup hand-held displays (Figure 1). SSC San Diego engineers extended this foundation with a commercial global positioning system (GPS) unit for location tracking, a compass and tilt sensor for head tracking, a wireless communication subsystem, and an instrumented glove for gesture control of display features. Head tracking permitted the development

FIGURE 1. Based on a commercial wearable computing system and augmented with SSC San Diego-developed software and display concepts, AITS is used for field surveillance and monitoring. With use of see-through display components and head-tracking technology, symbology can appear superimposed over the environment.

of three distinct display modes based on the operator's gaze:

1. When the operator looks up—a raw information display from whichever sensor initiated an alert.
2. When the operator looks down—a geo-referenced map presentation, synchronized to the user's location.
3. When the operator scans the horizon—discrete target cues and supporting information about the detected intrusion.

AITS provides a practical augmented reality interface for field use and permits security operators to monitor their sensor suite while on the move. Internet protocol extensions, described above, support data sharing by multiple security operators in real time, using the continuously updated map display. The AITS interface introduces a range of new display, interaction, and tracking capabilities at relatively low cost; SSC San Diego developers are currently testing user response to these design features.

Virtual Reality Technical Data System (VRTDS)

VRTDS was initiated as a component of the Network-centric Q-70 program under the sponsorship of the Space and Naval Warfare Systems Command (SPAWAR). VRTDS built upon a technical foundation established by AITS and is intended to support a variety of mobile shipboard tasks. The VRTDS design approach involves a selectable range of sensors, displays, computing resources, and interaction tools, all placed on a foundation of wireless communications technologies (Figure 2). VRTDS can present information in a variety of formats and incorporates augmented reality concepts for selected applications. Because VRTDS relies on proper selection and configuration of commercial components, the interface can be tailored in cost and capability, and can grow with new technologies. VRTDS emphasizes situation awareness and ease of operation for faster response and reduced training requirements.

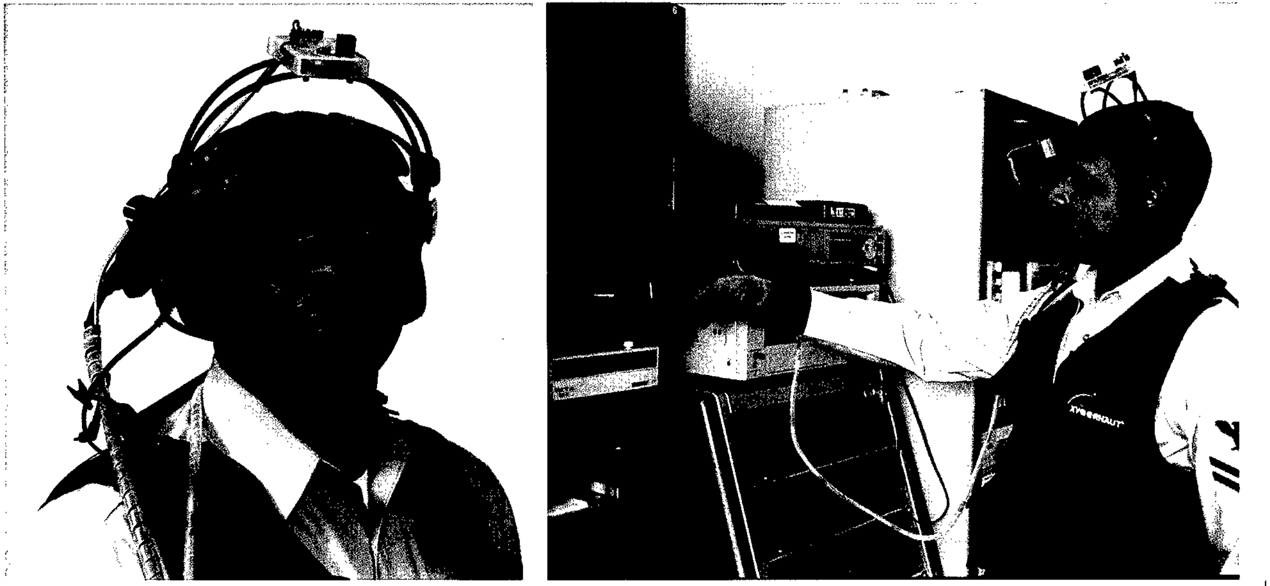


FIGURE 2. The VRTDS employs a see-through display concept, with graphics and text superimposed over the environment, which can provide maintenance and troubleshooting information directly in the user's field of vision. VRTDS displays can be controlled with gestures, using a specially instrumented glove.

The VRTDS development process is characterized by early and frequent involvement of operational communities (e.g., tactical decision-makers, maintenance personnel, and technical experts) concerning design features and functions. VRTDS has given explicit priority to information display and decision support issues, with technology selection and integration used only to realize a required information need. Shipboard functions targeted for VRTDS support include maintenance, emergency response, telemedicine, and command and control.

Maintenance

The visualization tools for maintenance support typically provide for the electronic display of equipment diagrams and text material. More sophisticated methods, however, can furnish the technician with views of the inner assemblies of equipment before maintenance begins. Such tools can also present amplifying information about equipment without making a person stop and consult manuals. Portable computing systems with flexible commercial software can even be employed in place of current test equipment, i.e., "virtual test instruments," providing both the computer processing and the visual interface for a variety of troubleshooting functions now supported by special-purpose devices. When maintenance tasks are completed, these same portable tools can be used to document the actions performed, the parts used or ordered, and the results of the repair effort—information that can then be uploaded to remote databases to support quality-assurance measures, trend analyses, material resupply, and scheduling of future tasks. Finally, advanced visualization and computing tools can be used to deliver maintenance training and procedures practice in order to keep seldom-used or complex skills sharp while deployed.

Emergency Response

Current damage-control activities are still coordinated almost exclusively with verbal communications. Data visualization using portable sensors, personnel tracking, and wireless communications tools can, however, disseminate a large volume of status information accurately and quickly to team leaders and to the ship captain in order to enable more rapid selection and efficient deployment of response resources. Expanded use of such tracking technologies can support real-time location of all personnel deployed in ship spaces, as well as report on their condition and welfare (e.g., through physiological and environmental sensors), thus greatly reducing the time required to locate and account for ship crew members during emergencies.

Telemedicine

It is a relatively straightforward matter to extend the application of maintenance and emergency response features, described above, to the needs of telemedicine. A combination of special sensors (e.g., physiological monitors, thermal and conventional imaging cameras), virtual test equipment concepts (to process sensor signals), on-site data stores, and wireless data sharing provide a complete foundation for mobile medical personnel to gather and transmit casualty data from the encounter site, to confer with remote experts, and to record care procedures for patient processing.

Command and Control

Finally, VRTDS components are being examined as interfaces for Navy command and control applications. Such interfaces could provide tactical information to the warfare commander without the space and power requirements of current workstation displays. Furthermore, this information would be available regardless of where the commander was physically situated in the ship. Such a distributed computing and visualization capability could, for example, permit personnel to monitor and control ship systems, evaluate tactical displays, and control weapons entirely from a variety of locations. Control authority is, however, a central issue beyond the realm of technology support; this application is, therefore, only exploratory.

SUMMARY

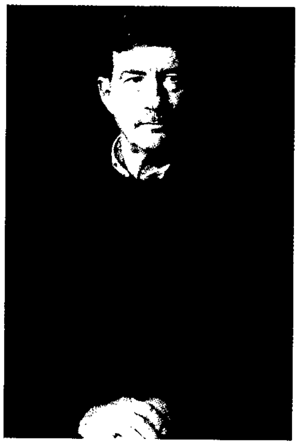
Wearable computing, portable visualization tools, and distributed communications tools have already proven their value for many shipboard activities; mobile information support, wearable computing systems, and wireless communications have all been successfully tested both ashore and aboard ship with the help of SSC San Diego engineers. Current SSC San Diego efforts are focused on incorporating additional government and commercial technologies into these mobile information systems, developing a stable testing facility, and coordinating efforts with other agencies.

Military and engineering leaders should be prepared to expect powerful new tools from these technologies and should also be prepared to think boldly when formulating management schemes to use such capabilities. In whatever form such systems evolve, however, SSC San Diego will have an important role to play to ensure that the Fleet obtains maximum benefit from its investment.

BIBLIOGRAPHY

1. Azuma, R. T. 1997. "A Survey of Augmented Reality," *Presence*, vol. 6, no. 4, pp. 355–385.
2. Murray, S. A. 1999. "Advanced Interface for Tactical Security (AITS): Problem Analysis and Concept Definition," TR 1807 (December), SSC San Diego, San Diego, CA.
3. Murray, S. A., D. W. Gage, J. P. Bott, D. W. Murphy, W. D. Bryan, S. W. Martin, and H. G. Nguyen. 1998. "Advanced User Interface Design and Advanced Internetting for Tactical Security Systems," 14th Annual National Defense Industrial Association (NDIA) Security Technology Symposium and Exhibition, 15 to 18 June, Williamsburg, VA.
4. Starner, T., S. Mann, B. Rhodes, J. Levine, J. Healey, D. Kirsch, R. W. Picard, and A. Pentland. 1997. "Augmented Reality through Wearable Computing," *Presence*, vol. 6, no. 4, pp. 386–398.

❖



Steve Murray

Ph.D. in Industrial Engineering, University of Wisconsin, Madison, 1997

Current Research: Augmented reality; wearable computing; operator alertness and performance measurement.

Adaptive Intelligent Agents: Human–Computer Collaboration in Command and Control Application Environments

Brenda Joy Powers
SSC San Diego

ABSTRACT

In the past decade, intelligent agents have proven to be of interest in many important application areas, such as electronic commerce on the Internet, the control of space probes on missions to the outer planets, the design of user interfaces, and military mission planning and execution operations involving decision-making and co-ordination functions—collectively known as command and control (C^2). C^2 application environments are dynamic and non-deterministic; thus, there are unique challenges involved in incorporating intelligent-agent technology within them. Decision-makers are required to assess and solve a variety of problems as quickly as possible, at times without adequate resources. The incorporation of agent technology into C^2 applications offers great benefit in the form of human–computer collaboration and provides decision-makers with assistance in carrying out their mission-related activities. This paper presents some suggestions on the types of tasks best suited to agents used in C^2 application environments and discusses the challenges involved in using agent technology within C^2 application environments.

INTRODUCTION

Command and control (C^2) application environments are characterized by their uncertainty and dynamism. This presents several challenges in implementing agent technology into them. Agents must be able to adapt to the changing circumstances and events of a military contingency, which means they must remain somewhat autonomous if they are to effectively assist human decision-makers in accomplishing their C^2 mission-related activities. Agents must possess enough autonomy to behave proactively in order to be of maximum benefit in a human–computer partnership. While this is true, the abilities of human decision-makers in the areas of conceptualization, abstraction, and creativity [1] far surpass their agent counterparts, whose strengths lie in computational speed, parallelism, accuracy, and data assimilation and management. Given these facts, this paper attempts to answer the following questions: (1) How can we effectively use agents to assist military decision-makers? (2) To what level can agents remain truly autonomous when humans must be kept in the loop? (3) Are there certain tasks that are better suited for agents to perform in C^2 application domains?

DEFINITIONS

This section defines some of the terms that will be used throughout this paper.

Autonomous Agents: Software and robotic entities capable of independent action in open, unpredictable environments. Autonomy has most often been defined as freedom from human intervention, oversight, or control [2].

Software Agents: Autonomous software entities that perform tasks on behalf of a user or another agent. Autonomous entities can assist users when performing their operations, collaborate with each other to jointly solve different problems, and answer users' needs [3].

Adaptive Agents: Webster's dictionary [4] defines "adapt" as the capability "to adjust (oneself) to new or changed circumstances." An adaptive agent can acquire knowledge (learn) and adapt (adjust) its behavior accordingly.

Multi-agent Systems: Multi-agent systems may be regarded as a group of intelligent entities called agents, interacting with one another to collectively achieve their goals [5]. Multi-agent systems implement distributed problem-solving, which provides many advantages including fast, parallel computing and increased fault tolerance [6].

Command and Control: Decision-making and coordination activities performed by military decision-makers during a contingency.

Human–Computer Collaboration: The ability of humans and computers to work together to solve problems. Specifically, while engaged in problem-solving and decision-making, humans contribute the ability to draw upon personal experience and intuition, and autonomous agents assist humans by providing superior speed, accuracy, and computational power.

AUTONOMOUS AGENTS IN C^2 APPLICATION ENVIRONMENTS

This section is divided into two parts. The first part gives an overview of current C^2 operations. The second part presents a domain example describing possible tasks that could be assigned to agents acting autonomously to assist decision-makers in accomplishing their mission-related activities.

C^2 Overview

The need for automating methods of accomplishing military C^2 activities is of utmost importance in today's military mission planning and execution operations. As previously defined, C^2 activities are those decision-making and coordination activities performed by military decision-makers. In combat, effective C^2 and success in battle requires commanders to develop associations and thought patterns. During a contingency, military commanders and their staffs must make timely and effective decisions under pressure. They often spend too much time manipulating information systems to filter data into meaningful information and performing routine tasks to assess the situation. It takes years of training and experience to develop the required skills to manage the pre-planning and subsequent engagement during a tactical encounter. Thus, even with advances in the area of intelligent systems, in C^2 environments humans must be kept in the "loop." Currently, most military C^2 activities performed by decision-makers are accomplished via paper and voice circuits. Toward this end, technology based on intelligent agents acting autonomously to perform user-specified tasks offers potential for automating and speeding up many of these time-critical activities. The next section focuses on human–computer collaboration within the context of a specific C^2 application domain example.

Domain Example

Air Warfare Operational Overview

Air warfare is defined in Joint Department of Defense publications as "the detection, tracking, destruction, or neutralization of enemy air platforms and airborne weapons, whether launched by the enemy from air, surface, subsurface, or land platforms." In an air warfare mission, the Air Warfare Commander (AWC), also known as the Area Air Defense Commander (AADC) for joint operations, is responsible for the development and distribution of an Area Air Defense Plan (AADP). The AADP, which contains the campaign plan and pre-planned responses used in dealing with the enemy air threat, is sent via teletype as a standard formatted military message called the Operational Tasks (OPTASK) Air Defense (AD), to all of the commanders in the battle group and subordinate air defense units, both afloat and ashore. The other significant report promulgated throughout the battle group is the OPTASK Link, which specifies the data link (communication) procedures within the battle group. Upon receipt, the individual commanders analyze the OPTASK AD and Link and generate plans for their respective region/sector of concern within the area of operations. Air defense planning also involves the coordination of air, surface, and mobile air defense assets. Decision-makers coordinate the allocation of scarce resources (airplanes, pilots, missiles, etc.) and work to minimize conflicts between competing engagements. This process is known as maintaining situational awareness. One of the main objectives of the AWC/AADC and his subordinates during the contingency is to maintain situational awareness. Table 1 lists the information they must keep track of in order to accomplish this objective.

TABLE 1. Situational awareness description.

Enemy
Locations (latitude-longitude, grid position, etc.)
Resources (troops, aircraft, tanks, artillery, etc.)
Status (in garrison, deployed, etc.)
Possible actions (attack, defend, reinforce, withdraw)
Friendly
Locations
Resources (platforms)
Status (combat ready, deployed, inside the continental U.S. [INCONUS], etc.)
Control measures (fire support coordination lines, restricted fire areas, phase lines, etc.)
Planned actions (e.g., OPTASK AD, pre-planned responses, etc.)
Logistics (Friendly and Enemy)
Locations
Resources (fuel, ammunition, food)

The report generated in conjunction with maintaining situational awareness is called a situation report (SITREP). Currently, this is a voice report that is required once an hour from all warfare commanders in the battle group.

The next section presents suggestions about opportunities for human–computer collaboration in a Littoral Air Defense mission. Some ways that autonomous agents can assist decision-makers in carrying out C^2 activities, such as formulating pre-planned responses and maintaining situational awareness, are discussed.

Agents in a Littoral Air Defense Environment

Picture a littoral air defense environment (operating close to the shore), where the Joint Forces Air Component Commander (JFACC) is responsible for coordinating theatre air defense among Joint and Allied forces. U.S. forces are involved in a major regional contingency located off the coast of California. The commander responsible for air defense is the Area Air Defense Commander, and is located ashore in an underground command center collocated with the Combined Forces/Joint Task Force Commander. Now we consider some of the specific tasks that agents could be assigned to assist decision-makers in the context of a littoral air defense mission. The AADC's first task will be the formulation of the pre-planned responses contained in the OPTASK AD. To accomplish this, the geographical constraints of the battle space and the evaluation of the enemy and assessment of its capabilities must be considered. The constraints of geography in the battle space must be considered because the contingency is located in confined waters. The battle space may be defined as a conceptual bubble around a friendly force in which a commander feels comfortable in detecting, tracking, and engaging threats before they can pose a significant danger to his vital units/defended asset list. Assume the commander is also constrained by physical "borders," such as reefs or shallows, or territorial borders such as the 12-mile limit, in the positioning of surface-to-air missile picket ships or screening platforms. These factors further reduce the reaction time allotted to any threat that does materialize. Agents with expert knowledge of the specifics of the topology of this region could take the initiative, generate potential plans for attack/ defense, and present them to human decision-makers for acceptance or rejection. Another task that must be accomplished is the generation of the OPTASK Link message. Currently, the OPTASK Link report is prepared manually, using a chart and cross-referencing the communication protocols for each asset in the battle group to come up with the list of who can talk to whom. Clearly, this is a cumbersome task that could be automatically handled by an agent that could simply retrieve the necessary information, cross-reference it, and produce a report in a fraction of the time. Upon completion, the agent could present the OPTASK Link to the user for transmission.

Some tasks that agents could perform to help decision-makers maintain situational awareness include keeping track of both friendly and enemy logistics (see Table 1) and monitoring weather conditions. For example, an agent might be assigned the task of keeping track of how many missiles the enemy has. Agents that have access to knowledge about enemy order of battle, (the list of enemy assets) could recommend the optimum shot and determine vulnerabilities. Weather data should be updated periodically, a task that could be performed by a monitoring agent assigned to that particular type of information. For example, if an agent detects an approaching storm, it would then know to advise the decision-maker to suspend air operations temporarily. The agent would also check to ensure that the ship's fuel level was not less than 50%. If the fuel level was less than 50%, action would need to be taken. Fuel level seems like a small detail, but the consequences of a ship running out of fuel and not being able to refuel could be disastrous. Consider that decision-makers are already under a large amount of stress in a contingency, and that declarative memory power is reduced in such a situation. The commander has already been advised to know the enemy capabilities, which involves the analysis of all the ships, aircraft, and submarines that could be encountered. Clearly, this is not a trivial task because it involves the ability to commit a large amount of information to memory. Agents with expert knowledge can provide platform-specific guidance when the need arises, thereby reducing the chances of error in decision-making. There is no reason why a decision-maker should have to keep track of and remember these kinds of details when agents, which are independent of reactions to stress, can assist.

CONCLUSION

The need for automating methods of accomplishing military C^2 activities is critical in today's military mission planning and execution operations. This paper presented some suggestions on the types of tasks autonomous agents operating within C^2 application environments could best perform. These tasks could best be performed in a littoral air defense environment and include assisting decision-makers in maintaining situational awareness, keeping track of both friendly and enemy logistics, monitoring weather conditions, providing information about the geographical constraints of the battle space, and gathering data on the communication protocols for each asset in the battle space and producing a report. Agents need to maintain a minimal degree of autonomy to be of maximum use to decision-makers involved in performing their mission-related C^2 activities. For example, agents, unlike human decision-makers, can keep track of vast amounts of information and do not experience stress in crisis situations. Thus, agents with expert knowledge of enemy capabilities and enough autonomy to determine a need for action could provide platform-specific guidance, thereby reducing the chances of errors in decision-making.

Future research is required to establish the degree to which agents should remain autonomous when acting as planning and decision aids for military decision-makers. Additional research is also needed to prove that the tasks identified in this paper are the types of tasks best suited to agents operating in C^2 application environments.

ACKNOWLEDGMENT

The SSC San Diego In-house Laboratory Independent Research (ILIR) Program funded this effort.

REFERENCES

1. Pohl, J., A. Chapman, and K. Pohl. 2000. "Computer-Aided Design Systems for the 21st Century: Some Design Guidelines," Fifth International Conference on Design and Decision-Support Systems for Architecture and Urban Planning, 22 to 25 August, Nijkerk, The Netherlands.
2. Barber, K. S. and C. E. Martin. 1999. "Agent Autonomy: Specification, Measurement, and Dynamic Adjustment," Technical Report TR99-UT-LIPS-AGENTS-09, (May), University of Texas at Austin, Austin, TX.
3. Maamar, Z., N. Troudi, and P. Rostal. 2000. "Software Agents for Workflows Support," *Journal of Conceptual Modeling*, no. 12 (February).
4. *Webster's New World Dictionary*. 1991. Simon and Schuster, New York, NY.
5. Barber, K. S., T. H. Liu, and D. C. Han. 1999. "Agent-Oriented Design," *Multi-Agent System Engineering: Proceedings of the 9th European Workshop on Modeling Autonomous Agents in a Multi-Agent World*, MAAMAW '99, 30 June to 2 July, Valencia, Spain.
6. Barber, K. S., T. H. Liu. 2000. "Conflict Detection during Plan Integration Based on the Extended PERT Diagram," Fourth International Conference on Autonomous Agents (Agents 2000), 3 to 7 June, Barcelona, Catalonia, Spain.

❖

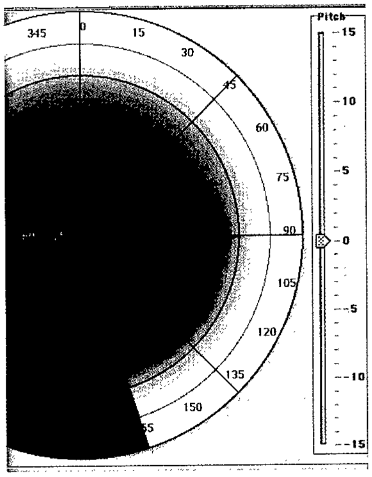
Brenda Joy Powers

BS in Computer Science, Point Loma Nazarene University, 1984

Current Research: Intelligent agent technology; object technology.

5

Communication Systems Technologies ■



5 COMMUNICATION SYSTEMS TECHNOLOGIES

Strategies for Optimizing Bandwidth Efficiency

Todd Landers
SSC San Diego

ABSTRACT

To optimize bandwidth efficiency, the natural limitations of each network-supported data type must be overcome or mitigated. This paper discusses issues affecting bandwidth efficiency through the U.S. Navy's bandwidth-constrained wide-area network (WAN). The paper details the prevalent data types found in the naval environment and describes the characteristics associated with each data type. Commercial, standards-based link layer protocols that have widespread application in Navy networks are also described. Finally, forward error correction and issues surrounding bandwidth efficiency are discussed.

DATA CHARACTERISTICS

Loss-Sensitive Information

The data type envisioned when discussing data communications is usually loss-sensitive data. Many network engineers mistakenly assume that loss-sensitive data are the predominant data type on the Navy's general-purpose wide-area networks (WANs). This data type must be faithfully reproduced with 100% accuracy at the distant end of the link before it can be used. The Transmission Control Protocol (TCP)/Internet Protocol (IP) usually transmits the data type as computer data/software. Any differences in the data reproduced on the distant end of a link make the entire communication unusable. If data are corrupt, the application of the data will be invalid. The time it takes the data to reach the distant end of the link has little effect on data usability. This data type is loss intolerant, but latency tolerant.

As mentioned, the bulk of this data type is transferred using the IP family of protocols. IP is inherently connectionless. It uses a 32-bit address scheme to identify hosts on the network. The User Datagram Protocol (UDP) and TCP use IP as its transport layer.

For broadcast applications, UDP uses the connectionless properties of IP to its advantage. It uses a checksum to verify data integrity and discards corrupt data. UDP is ideal for applications that are loss and latency tolerant, and can be used in those instances to help minimize unneeded router chatter over bandwidth-constrained links.

TCP adds a connection-aware element on top of IP. TCP has embedded mechanisms that check the content and sequence of arriving packets. TCP also allows hosts to set timers. The host may "time out" a connection, enabling the host to free up system resources that would otherwise be tied up with a suspected dead connection. TCP will automatically request a re-send from the originating host if loss or corruption is detected. TCP uses these and other tools to keep the network working smoothly as long as latency is managed at lower levels of the network.

Unfortunately, the same features that help TCP work well in situations where bandwidth is ample can be disastrous when bandwidth becomes constrained. Once a link in the network becomes bandwidth-constrained, the applications start asking for retransmission of data assumed lost (in this case, just delayed). This unnecessary request for information is the beginning of the end of data transfer across the link. If Ethernet is used as the link layer protocol, connection timeouts caused by long round-trip time can cause a router to start seeking alternate paths to the desired host. The additional router chatter contributes to the congestion of the already congested link. This congestion is a death spiral for a TCP connection. The connection is terminated, and if the host is looking for the original information, it attempts to reconnect. Note that no useful information is exchanged, though bandwidth is consumed. The bandwidth consumption prevents other worthy circuits from exchanging useful information.

The way applications use IP may cause other inefficiencies. If the payload of the IP packet is not appropriately sized for the data type conveyed, huge amounts of bandwidth could be consumed because bit stuffing is needed to make complete packets. The data type usually originates at hosts that provide sensor inputs (like voice) to another application. If a sensor needs to transmit a sample containing a few bytes to a remote host through TCP, the originator usually stuffs filler bytes into a packet with a length that is probably several hundred bytes. The efficiency of this connection quickly approaches zero, which is not a problem until bandwidth becomes limited. File compression can reduce the size of an application data file or sensor output before transmission through the WAN. File compression engines must be deployed to all source and user sites, which causes a logistics problem, but the overall gain in bandwidth efficiency is worth the trouble. File compression can reduce the actual data transmitted across the link by 80%.

Header compression techniques can reduce bandwidth consumed over a point-to-point link through various network protocols. For header compression to add value to the WAN, it must add minimal latency due to processing overhead and be completely symmetrical. Compression abbreviates redundant header information in a data stream before transmission over the WAN and then restores the header to its original state after it reaches its final destination. The higher the compression engine is in the protocol stack, the more opportunity to save bandwidth. The more aggressive the compression engine, however, the more latency it adds to the circuit. Header compression at the transport and network layers depends on some error checking at the link layer to be effective.

Time-Sensitive, Loss-Tolerant Information (Voice)

The most common type of time-sensitive, loss-tolerant data is plain old telephone system (POTS). Unlike computer-oriented information, the interpretive device for POTS is the human ear. Studies show that over 50% of a speech signal can be removed and the human ear can still assemble the required information to extract the audible message. However, as little as a 0.5-second delay can cause severe degradation of the intended communication. A system designed to handle large quantities of this data type can lose a lot of data, but if data are delayed or delivered out of order, it is useless, and interpreted as noise.

Networks specializing in this data type are quite different from those that handle large quantities of loss-sensitive data. In terrestrial networks where bandwidth is ample, a typical voice call is digitized and transmitted using a G.711 protocol through the public switched telephone network (PSTN) at 64 kbps. Toll-quality voice has an upper latency limit of a 200-ms delay across the network. These networks are composed of various sizes of public branch exchange (PBX) switches. The interconnections between PBXs generally scale in 64-Kbps chunks.

PSTNs are connection-oriented. When a call initiates, the originator transmits a setup preamble that negotiates for a connection at each intermediate switch along the way. If the connection cannot be supported at any point along the way, the entire connection is denied. If all attempts to establish the end-to-end connection are denied, the originator gets a busy signal. For most PSTN users, this busy signal only happens on Mother's Day or after a natural disaster. There is no such thing as a lower grade of service; a connection exists or it does not. During the call, a near-real-time connection for N x 64 kbps allows the user to talk, send a facsimile (FAX), or use a modem, secure telephone unit (STU), or secure telephone equipment (STE), etc. After the call is completed, a teardown sequence allows each switch in the circuit to release the resources reserved for that connection.

This type of network has many sources of inefficiency. First, voice is the most common type of connection supported through this network. Voice typically has less than a 50% duty cycle. Generally, only one person talks at a time, and usually there is silence between words. Everything else is dead air (wasted bandwidth). Silence-suppression techniques reduce the dead-air bandwidth consumption to help solve this problem.

Another source of inefficiency is that 64 kbps is not really needed to digitize and communicate using voice. The 64-kbps convention was adopted because it was an easily implemented solution, not because it was the most efficient. Several voice compression algorithms can drastically reduce the amount of bandwidth used for each voice call. Toll-quality voice has been compressed to 8-kbps or one-eighth of the bandwidth allotted for a typical voice call. Good-quality voice has been transmitted using less than 800 bps. Unfortunately, other applications using the PSTN do not respond well when compressed with some of the more aggressive compression techniques, so compression must be applied selectively.

Modems and FAX machines use the PSTN to transmit analog-modulated digital signals. This inefficient means of digital data transfer was developed years ago to overcome noisy analog transmission lines that were once used to interconnect PBXs and end-users. Connections have improved, but the format is outdated. The most efficient way to accommodate these types of connections is to convert the modulated digital signal back into ones and zeros and transmit them through the network using much less bandwidth. A FAX machine can be supported at 9.6 to 14.4 kbps instead of consuming the full 64 kbps allocated to each connection by the PSTN. STU-III can also be supported using this type of compression technique, but the modulation scheme must be implemented in accordance with National Security Agency (NSA) policy.

Time and Loss-Sensitive Information

The synchronous serial data type is traditionally used where the system designer had creative control of the entire system. These links are susceptible to loss of content and fluctuations in the end-to-end timing. Each communication link was usually built to support one application set. Interoperability and flexibility were not considered in the design. These systems are probably the single largest source of wasted bandwidth. After the communication link initiates, it remains active regardless of use.

Circuits had to be provisioned to support worst-case bandwidth needs, and as applications became more bandwidth-efficient, their bandwidth usage remained high and constant.

As networked applications became more popular, synchronous serial communications became known as communications "stovepipes." The Department of Defense has invested huge amounts of resources into developing and refining stovepipe systems over the past 30 years. Although new systems focus more on networked solutions, stovepipes are still with us today primarily because of the cryptography developed to support legacy applications. The slow development of network cryptography has hindered application development and subsequent migration away from stovepipes.

Video

Video is generally more tolerant of timing than synchronous serial connections, but jitter is deadly. The type of video compression used should vary depending on the video content. Compressed video usually transmits all information needed to paint the screen the first time, and transmits only the changes to the initial image. Regardless of the resolution or quality of the video, this approach allows video to be supported using variable bit-rate service contracts through the network. Talking-head videoteleconference (VTC) video should use the most aggressive video compression techniques. This type of application can operate well on less than 64 kbps.

One of the worst misuses of bandwidth for video traffic occurs when the host or the network provisioning creates a fixed-bandwidth pipe for the video call. H.320 is a common video compression format used with Integrated Services Digital Network (ISDN) networks. Each video call allocates N x 64 kbps to support the call resolution selected by the user. The bandwidth within the fixed allocation continues to fluctuate; however, even though the bandwidth need reduces when the picture stabilizes, no bandwidth is available for other applications.

Although still both time and loss sensitive, the H.323 protocol is much more tolerant in both areas. H.323 works with IP networks and has progressed in overcoming some of the inherent obstacles for supporting voice and video on an IP network. Unfortunately, H.323 is susceptible to many problems that plague data transmission over IP-based WANs. While some jitter or delay can be tolerated, excessive congestion can cause the H.323 session to freeze. Though there is some time-sensitivity, packets are mixed with and sometimes delayed because of packets that have no time-sensitivity. To overcome this deficiency, priority queuing can reduce the likelihood that the time-sensitive video will incur fatal transmission delay.

The screen capture in Figure 1 shows the bandwidth consumption of the H.323 protocol generated using Microsoft Netmeeting. The link information shown represents one-half of a bidirectional link. The video resolution for this example was set up to run at best-fidelity voice and video. H.323 is highly variable in its bandwidth requirement (red trace in Figure 1). The

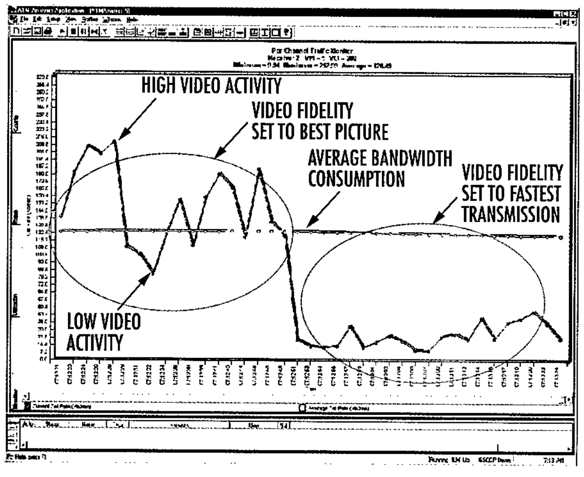

HIGH VIDEO ACTIVITY
VIDEO FIDELITY SET TO BEST PICTURE
AVERAGE BANDWIDTH CONSUMPTION
VIDEO FIDELITY SET TO FASTEST TRANSMISSION
LOW VIDEO ACTIVITY

FIGURE 1. Bandwidth consumption of the H.323 protocol.

surges in bandwidth usage occur when the video compression engine must transmit updates to large portions of the image. With video quality set to its highest setting, the average bandwidth usage settles out near 130 kbps, with surges up to nearly 190 kbps. During periods when the video does not change, bandwidth usage drops to below 80 kbps. Netmeeting allows the user to reduce the fidelity of the picture to accommodate bandwidth-constrained connections. At the lowest resolution settings, the average bandwidth usage has been observed below 20 kbps for a full-motion VTC.

TYPICAL NAVY WAN DATA LOADING

The U.S. Navy operates in a truly converged WAN environment. Figure 2 shows a sample of the circuits assigned to USS *Coronado* (AGF 11), a command ship, during a typical deployment. The bandwidth available on *Coronado* should be considered the best possible case because she has been outfitted with the best communications available in the U.S. Navy to support various developmental enterprises.

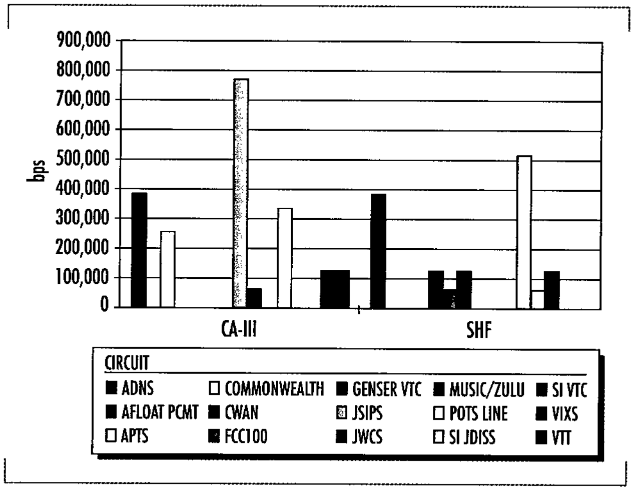


FIGURE 2. Circuits assigned to *Coronado* during typical deployment.

Coronado runs multiple T-1s using various super high-frequency (SHF) and Challenge Athena configurations.

Voice, video, and data must all co-exist on the WAN (Figure 2). The major users include a Joint Service Imagery Processing System (JSIPS), which is an intelligence circuit currently using a synchronous serial EIA-530-based system. This circuit was one of the primary reasons for the procurement of the Challenge Athena system, so when this circuit becomes active, other lower priority circuits are manually disconnected. The JSIPS circuit is a prime example of current bandwidth management practices. Other large data users include secure and non-secure voice (both circuits are listed in the figure as POTS LINE). POTS lines are supported using compressed voice cards in various time-division multiplexers (TDMs). The compression cards reduce the bandwidth required to support each voice call from 64 kbps to between 8 and 16 kbps. The compressed voice signals are aggregated as synchronous serial circuits before porting to the satellite communications (SATCOM) modems.

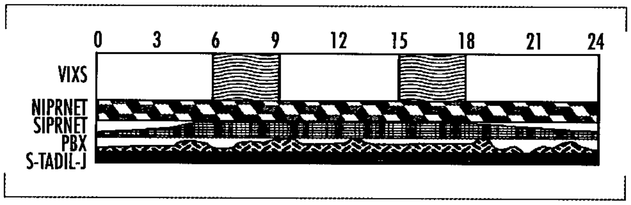


FIGURE 3. TDM circuit.

In the bandwidth management approach, fixed-bandwidth synchronous serial circuits constrain circuits that use protocols that dynamically consume bandwidth such as IP. The circuits marked "ADNS" (Automated Digital Network System) represent the wide-area IP-based traffic, and typically are assigned up to 384 kbps, supporting a mixture of classified and unclassified data.

Figure 3 shows how the circuits on the Y-axis might consume bandwidth when provisioned through a TDM using today's provisioning approach. The white space represents provisioned, but unused, bandwidth. A few points to notice in this figure are as follows:

- Wasted bandwidth by low-usage, high-bandwidth systems such as the Video Information Exchange Subsystem (VIXS), which is a H.320-based VTC using synchronous serial cryptography for security

- Congestion in the low-priority, high-usage circuits such as NIPRNET (unclassified but sensitive IP router network)
- "Nailed up" circuits consuming bandwidth to keep the circuit timing alive, such as Satellite Tactical Data Information Link–Joint (S-TADIL–J)
- More bandwidth available to support additional voice and data circuits than are allowed under current provisioning policy

LINK-LAYER CONSIDERATIONS

Ethernet

Ethernet is commonly used as the link-layer protocol for TCP/IP and UDP/IP. Ethernet is a very cost-effective way to deliver data to end-users. The network equipment is inexpensive and mature; there is a large application base with drivers supporting Ethernet; and there is a large pool of competent network administrators who understand the technology. Ethernet is inexpensive to deploy and administer. It scales very easily to 10 Gbps on the backbone. Ethernet uses the Carrier Sense Multiple Access with Collision Detection (CSMA/CD) approach to sharing bandwidth among the network users. If there is a collision, it simply re-sends the information. Ethernet works best on generously provisioned networks; for the most part, the assumption of few collisions is true.

Ethernet starts having difficulty when congestion occurs. At approximately 40% of the rated network throughput, Ethernet begins to bog down. At 60% link saturation, the link becomes nearly unusable because much of the traffic is re-sent from prior collisions.

Switched Ethernet technology provides answers to some of these problems by explicitly controlling traffic destined for users. An Ethernet switch can support single dedicated or multiple users on a switched segment. The switch logically separates the segments to eliminate collisions between segments. Multiple user segments continue to have the contention problem among subnetwork users.

Ethernet is problematic in the wide area because the organization deploying the local network will not have control over congestion in the wide area. Any single link in the wide area will slow performance experienced by the user. Network performance is only as good as the slowest link in the WAN.

Point-to-Point Protocol

Point-to-point protocol (PPP) is an encapsulation approach to transmitting IP over serial point-to-point links. PPP is a very flexible approach to transmitting IP datagrams through a serial link. The only real limitation PPP imposes on the link is that it has to be a full-duplex link. It supports synchronous and asynchronous transmission. It works fine over a number of common physical media including EIA 530, RS-232, V.35, etc. However, PPP links can cause unwanted latency and jitter because of the variable nature of the IP datagram contained in the data payload of the PPP frame.

ATM

Asynchronous transfer mode (ATM) effectively transmits a wide variety of data across a network. The size of the ATM cell (53 bytes) was developed as a compromise between the voice camp (small, prompt data delivery) and the IP camp (large, continuous streams of guaranteed data delivery). While ATM was originally envisioned to work on high-speed networks (OC-3 and above), it has been adapted for the WAN because it works through congested links.

ATM statistically multiplexes fixed-size cells through a link. The protocol organizes cells into logical or virtual point-to-point circuits through an interface. At the time of circuit setup, each interface in the circuit establishes a service contract with its neighbors. Each switch has a unique address to ease automated connection setup. Once all of the interfaces in the path have established the required service contracts, the data transfer begins. Cells with that circuit identifier are automatically switched along its path to the end-user. Once the data transmission is complete, the contracts are canceled and the circuit is disestablished.

The service contracts have built-in quality-of-service features. At the top layer, there are ATM Adaptation Layers (AALs). Each AAL has some predetermined characteristics and some preconceived notions of what applications that adaptation layer would support. For example, AAL-1 supports synchronous serial connections and looks to users like a static TDM. AAL-2 supports voice, and while it maintains the time relationships between cells, it can take advantage of the other characteristics of voice discussed earlier. AAL-5 supports IP and has many features to take advantage of the characteristics of IP data transfer.

Forward Error Correction

Forward error correction (FEC) is commonly applied on noisy links to improve error performance and, thus, the performance of the link. The different FEC algorithms include block, convolution, and veterbi codes. For the purposes of this paper, FEC can be applied in varying degrees to reduce the error rate of a link; however, the more rigorously FEC is applied, the more bandwidth overhead and processing latency increases.

As discussed, different data types have varying degrees of error tolerance. FEC should be tailored to the data type passing through the link. For example, voice is loss tolerant and will perform well even when the link has some errors. Some synchronous data streams are very loss sensitive and may cause the end-user equipment to malfunction if there are too many errors on the link. TCP/IP may start flooding the link with re-sends if the error rate is too high, thus causing data congestion.

Recently, many products have been shipping with adaptive FEC. This approach samples the noise on the link and adjusts the FEC algorithm to keep the link error rate nearly constant. As the FEC is applied more aggressively, the effective throughput drops. Applying adaptive FEC at the link layer provides better performance by reducing the amount of re-sends by the hosted applications.

Link Design Decisions

As with all aspects of an engineered solution, engineers must chose the best tools to confront each aspect of the link design. Because of the economics and maturity of the technology, Ethernet is a clear choice in the local area networks, but falls short in the wide area. PPP is a good choice if all of the applications supported by the network are IP-based, but PPP falls short for a general-purpose network that supports various data types. With the technology currently available, ATM is the only technology discussed that can support a truly converged network supporting voice, video, data, and legacy applications. Figure 4 shows the dynamic bandwidth allocation achieved using ATM for the WAN.

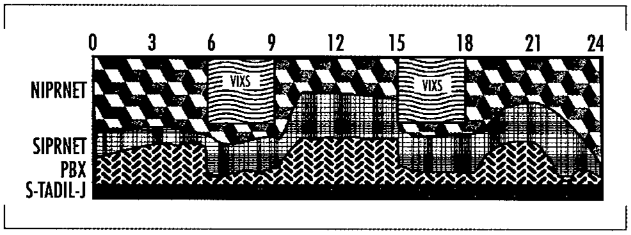


FIGURE 4. ATM circuit.

CONCLUSION

We can objectively maximize the information payload, optimizing for communication channel size, reducing non-productive information transfer, and using aggressive compression and forward error correction techniques. To get the greatest use from a SATCOM system, we must maximize the gross bandwidth efficiency. Additionally, user application data must be optimized. TCP/IP or UDP/IP solutions should be implemented wherever possible. Finally, the correct link layer technology must be selected for the environment in which it will be deployed.

Successful information transfer occurs only when enough data are transferred from the source through a data link to the distant end of the link to facilitate reassembly of the information suitable for end-device perception. Various types of information will be transmitted through our communications links. Some information is loss sensitive; other types of data are time sensitive; and still others are time and loss sensitive. They are all vital to the operation of the Fleet.

There are opportunities at every layer to optimize. The fiscal cost of not optimizing is tremendous. The operational cost could be devastating.

❖

Todd Landers
BS in Electrical Engineering, San Diego State University, 1985

Current Research: Wide-band ADNS architecture development; Tactical Switch System development; DoD Teleport requirements analysis and system definition; design/test of the IT-21 block one baseband system.

Tools for Analyzing and Describing the Impact of Superstructure Blockage on Availability in Shipboard and Submarine Satellite Communications Systems

Roy A. Axford, Jr.
SSC San Diego

Gerald B. Fitzgerald
The MITRE Corporation

ABSTRACT

The blockage analysis tools in the satellite communications (SATCOM) Availability Analyst (SA2) software package combine topside blockage data, communications satellite constellation positions, and ship-motion models to calculate the impact of superstructure blockage on availability as a function of antenna installation locations, ship's geographic position, and sea state. This impact can be evaluated for a specific position, along a ship's planned route, or averaged across the entire field of regard of the SATCOM constellation(s) of interest. This paper details the capabilities of the blockage analysis tools in SA2. The tools are applied to the analyses of topics of current interest including International Maritime Satellite (INMARSAT) on the CG 47 class, Global Broadcast Service (GBS) aboard the flagship USS Coronado *(AGF 11), and submarine High Data Rate (SubHDR) on the SSN 688 class.*

INTRODUCTION

On most of today's warships, it is impossible to find a single location for a satellite communications (SATCOM) antenna that provides an unobstructed view of the entire sky. If there is sufficient available topside space, two antennas are usually installed to support a mission-critical system (e.g., protected extremely high frequency [EHF] SATCOM). Lower priority systems are often forced to use a single antenna.

No matter how many antennas are used, it is critical to quantify the impact of topside blockage on the availability of a shipboard SATCOM system. Knowledge of this impact is needed in antenna location selection to ensure that the highest priority systems have the best views of the sky. Presenting this knowledge in an easily understood manner can make the topside design process more successful. Furthermore, once a shipboard antenna system is installed, the ship's company must have a clear understanding of the impact of unavoidable blockage on communications availability.

Many ship captains say that there are times when their choice of heading is dictated by whether or not a particular antenna system can "see" a desired satellite. It also appears that the determination of unblocked headings is often made by trial and error at sea, without benefit of *a priori* knowledge of the blockage situation of the SATCOM terminal in question. Topside blockage is so frequently discussed in the Fleet that there is widespread need for a software tool that can present this knowledge clearly.

This paper describes a set of tools for the analysis and display of the impact of superstructure blockage on shipboard SATCOM availability. These tools can give a ship's crew real-time indications of the blockage situation for an antenna system of interest with any desired geostationary satellite, based on the ship's present position, heading, and the Sea State. (Geosynchronous satellites in inclined orbits are discussed later in the Spatial Model section.) For route planning, there is also a display that shows the blockage situation along an entire Great Circle path as a function of Sea State (for the Great Circle headings). For more general planning and analysis, there is a display that shows the percentage of blockage-free headings (as a function of position and Sea State) as a colored cell on a global map. Along with their value to the operational community, these tools help topside designers compare the relative merits of candidate antenna installation locations.

The blockage analysis and display tools described here are components of a larger program called SATCOM Availability Analyst (SA2). The next section, Component Models, describes the lower-level components used to model the effects of blockage in SA2 (i.e., inputs). The Displays and Metrics section presents SA2's blockage analysis products (i.e., outputs). The Applications section gives some illustrative examples of SA2's recent application to the analysis of blockage for emerging SATCOM terminals to be installed on Aegis cruisers and *Los Angeles* class submarines.

COMPONENT MODELS

SA2 was developed as an extension of the Global Broadcast Service (GBS) Data Mapper (GDM) [1, 2]. GDM combined a simple Java-based Geographic Information System (GIS) with encodings of relevant International Telecommunications Union Radiocommunications (ITU-R) Recommendations and GBS link budget parameters to develop global maps of GBS link margin and availability.

Spatial Model

The core of SA2 is this same GIS, built upon a simple raster model of the earth's surface, which is represented as an array of 2.5° x 2.5° *model cells* (between latitudes 70S and 70N). The model-cell center points are stored as 3-space (x, y, z) vectors. The model-cell size can be varied, but 2.5° represents a good trade-off between precision and run time for most applications.

SATCOM constellations in SA2 are also represented as sets of 3-space vectors—each vector giving the Clarke Belt position (CBP) of one geosynchronous satellite. SA2 computes the elevation and azimuth angles from the center point of any model cell to a satellite's CBP by using simple vector arithmetic. These angles may then be combined with a ship's heading (entered by the SA2 user), and used as indices into a blockage matrix. As described in the following section, this matrix is an image of the superstructure blockage as seen from each antenna assigned to the satellite of interest. Thus, SA2 computes whether or not a ship in a given model cell and on a particular heading has an unblocked line of sight (LOS) to the satellite of interest. In addition, the SA2 user may enter a Sea State. SA2 then uses ship-class-dependent motion models (described later in the section on Ship Motion Models) to expand the LOS from the antenna into an appropriately distorted cone, thus accounting for the impact on satellite visibility. Increased ship motion in higher Sea States reduces availability by causing the superstructure to move in and out of an antenna's LOS to the satellite of interest. As shown in examples below, some antenna locations suffer more from this effect than others.

Many geosynchronous communications satellites of interest are in inclined orbits (e.g., the Ultra-High-Frequency [UHF] Follow-On [UFO]/Global Broadcast Service satellites: UFOs 8, 9, and 10). The pointing angles to such satellites from a geographic position vary over the diurnal cycle. SA2 does not model this motion, but it is able to read satellite track files in the form of pairs of azimuth and elevation pointing angles versus time for the position and satellite of interest. Such track files are readily available from applications such as Satellite Tool Kit (STK) and Satellite Orbit Analysis Program (SOAP). SA2 can use these pointing angles similarly to those for geostationary satellites (i.e., a geosynchronous satellite in an orbit with 0° of inclination with respect to the equatorial plane) to determine if the LOS is unblocked.

Blockage Models

Only the highest antenna on a ship can view the entire hemisphere of sky above it, and even then, only if the antenna is also the highest structure of any kind on the ship. Otherwise, additional antennas, masts, exhaust stacks, weapons, yardarms, or any other superstructure will mask out (i.e., block) some of the sky. The cluttered view of the sky from a shipboard SATCOM antenna's topside location is represented in SA2 as a two-dimensional, binary-valued matrix, with 360 columns covering azimuth angles in 1° increments, and 106 rows, covering elevation angles from –15° to the zenith (90°). (The zenith row is a degenerate case; all entries are identical.) Depression angles (elevation angles below 0°) must be included because as a ship rolls and pitches, the apparent elevation angle to a satellite near the horizon (relative to the ship's deck) may be negative.

Blockage matrices may be imported into SA2 by two methods. For installed terminals, the blockage information often already exists in a Blockage Adaptation Module (BAM) file generated from a digital image(s) of the view(s) from the antenna location(s). Alternatively, direct processing of such a digital image can create a blockage matrix for SA2. A suitable image could be acquired from the antenna's installation location with a fisheye-lens-equipped camera, but this method has not been used to obtain any of SA2's blockage data thus far. Almost all of the blockage data used in SA2 come from images generated by a three-dimensional computer-aided design (CAD) model of the ship's entire topside. Such topside models are often refined and/or updated by taking theodolite surveys of the ship's topside directly from the intended antenna installation location(s). (Reliance only on a ship's design drawings can lead to the omission of superstructure that was added after initial construction.) The image-processing software used is external to SA2. The software can acquire and digitize images in either polar or rectangular projections, and the software is an extension of a MITRE-developed image processing and exploitation suite originally written for the National Imagery and Mapping Agency (NIMA). Figure 1 presents a CAD-model topside blockage image typical of those that have supplied most of the blockage matrices available within SA2.

Note that these matrices model boresight or optical blockage. SA2 does not consider the near-field patterns of shipboard antennas or effects such as knife-edge diffraction. (This approach is supported for frequencies above ~1 GHz by conclusions of a study [3] in which detailed tests and analyses were performed to determine the blockage effects of various topside structures on the performance of the AN/USC-38 EHF shipboard SATCOM terminal.) However, antenna beamwidth can be simulated within SA2 by a simple, run-time operation that pads each blocked area by a user-specified number of degrees. A similar procedure is often used in producing the BAM files of

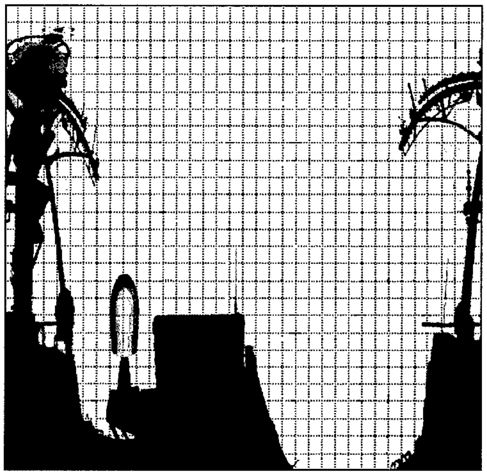


FIGURE 1. 3-D CAD topside model blockage image (rectangular projection) of the view from one of the INMARSAT antenna locations on the DDG 51 class.

dual-antenna systems to mark a "warning track" for the initiation of antenna handover procedures. Furthermore, BAM files often designate some of the smaller (say, less than 5 to 10 degrees in azimuth or elevation extent) unblocked areas as "blocked" to avoid "peep holes" that are lost due to ship's motion in moderate Sea States. This practice has also been adopted in producing the blockage matrices used SA2.

For dual-antenna systems, SA2 determines when an unblocked LOS to a desired satellite is available from *either* antenna or from *both* antennas. The "either antenna" mode analyzes systems that perform a hand-over from antenna "A" to antenna "B" as "A" moves into a "warning track." In these systems, either antenna can provide 100% of the communications services if it has an unblocked LOS to the satellite. The "both antennas" mode analyzes systems such as INMARSAT B High Speed Data (HSD) in which both antennas track the same satellite simultaneously to provide higher total throughput by using multiple transponder channels. In these systems, both antennas are required to provide 100% of the communications services.

Ship-Motion Models

As ships are accelerated by the wind and waves through which they travel, they experience Sea-State-dependent perturbations that are described by three rotational motions (pitch, yaw, and roll) and by three translational motions (surge, sway, and heave). These effects are detailed in [4]. Following McDonald's ranking of the magnitudes of these motions, SA2 confines itself to the impact of pitch and roll. Sea States high enough to make the other motions significant with respect to blockage are so severe that they surpass the operational specifications of Navy shipboard SATCOM antennas.

In [4], McDonald provides tables of length at the waterline, beam at the waterline, metacentric height and roll constants for various surface-ship classes. McDonald combines these constants with the ship-motion equations of DoD-STD-1399-301A [5] to provide ship-class-dependent pitch and roll extremes and periods as functions of Sea State. The resulting sinusoidal ship-motion equations are used to produce pitch and roll angles as functions of time for an animated display in SA2 in which the observer's frame of reference is the ship. These equations also allow the computation of temporal statistics (e.g., unblocked time/blocked time, durations of blocked times, etc.) that are of potential value in the evaluation of protocols for intermittent links.

Motion data are not as readily available for submarines. Since a helmsman actively controls the pitch of a submarine at periscope depth by using the stern planes, pitch is not a key factor in determining LOS availability at moderate Sea States. Roll is important, however. Submarine roll rates (even more than surface-ship roll rates), depend on the heading of the boat relative to the swell direction. For analyses of blockage aboard submarines, we have relied on interviews with former submariners to characterize the expected pitch-and-roll extremes and periods.

DISPLAYS AND METRICS

The data from the models described above are used in SA2 to provide blockage information for a ship's position along a route of travel or averaged over the entire field of regard of a satellite constellation of interest.

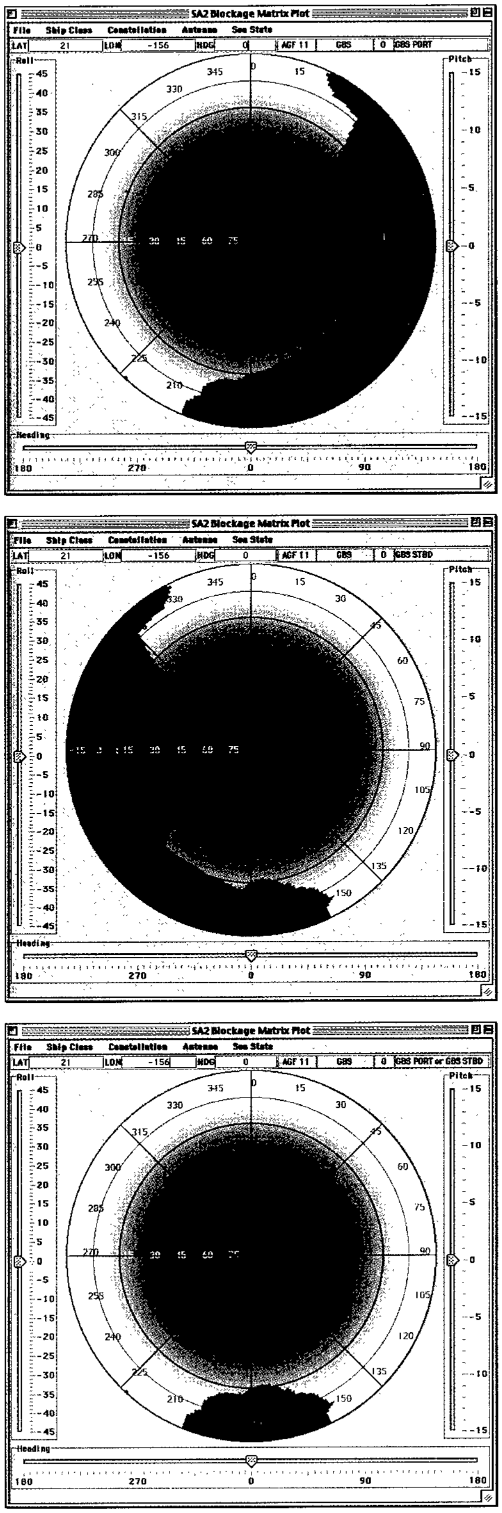
Ship's Position Blockage Information

SA2 provides an extremely useful display to assess the availability of a given SATCOM system from the ship's current position (or any position of interest) as determined by superstructure blockage and Sea State. Figure 2 provides three examples of SA2's SkyView display [6]. The currently selected blockage matrix (e.g., for a single antenna or for the composite blockage of two antennas) is always displayed in a polar orthographic projection. Ship's position may be typed in or entered by clicking on the desired spot on SA2's map display (see Figure 3). Ship's heading may also be typed in or adjusted with a slider. The satellites of the selected constellation (those above the horizon for the entered position) are then plotted as an overlay at the azimuth and elevation pointing angles computed for calm seas (i.e., for a level ship). Blocked satellites are shown in red, and visible ones are shown in green.

As noted above, the user can also examine the effects of Sea State with the SkyView display. Sliders allow the user to enter static pitch-and-roll angles based on, for example, the way the ship is behaving while underway (or to account for a list at the pier). The satellite's "dot" moves accordingly and turns red if it moves into a blocked region. Alternatively, using the pitch-and-roll magnitudes from [4], SA2 will plot the entire ship's motion envelope for a user-entered Sea State, resulting in green, red, or green and red "satellite smears" over the extent of pitch and roll. (Obviously, the user must correctly enter the ship's class for this approach to be useful.) The full equations of motion (according to the ship's class) may also provide an enlightening real-time animation of the apparent satellite positions. For example, see Figure 2 and imagine the satellite position moving according to the ship's equations of motion. (It is actually possible to get seasick while watching this display!) As the satellite moves, it turns red when blocked and green when unblocked. With all of these approaches, if any red appears, it indicates that the ship's motion might be causing intermittent outages for the SATCOM system in question. Thus, the SA2 SkyView display provides an aid for troubleshooting at sea.

Ship's Route Blockage Information

The SkyView display can help analyze SATCOM availability at a moderate number of positions, but availability along an entire ship's route is best viewed on SA2's TrackView display. With the aid of a text editor, the user enters pairs of end-points that are then connected by SA2 through use of a Great Circle route. The resulting tracks are color-coded along their extent according to the availability of the currently selected satellite or constellation of satellites.



SA2 Blockage Matrix Plot
LAT 21 LON -156 HDG 0 AGF 11 GBS 0 GBS PORT

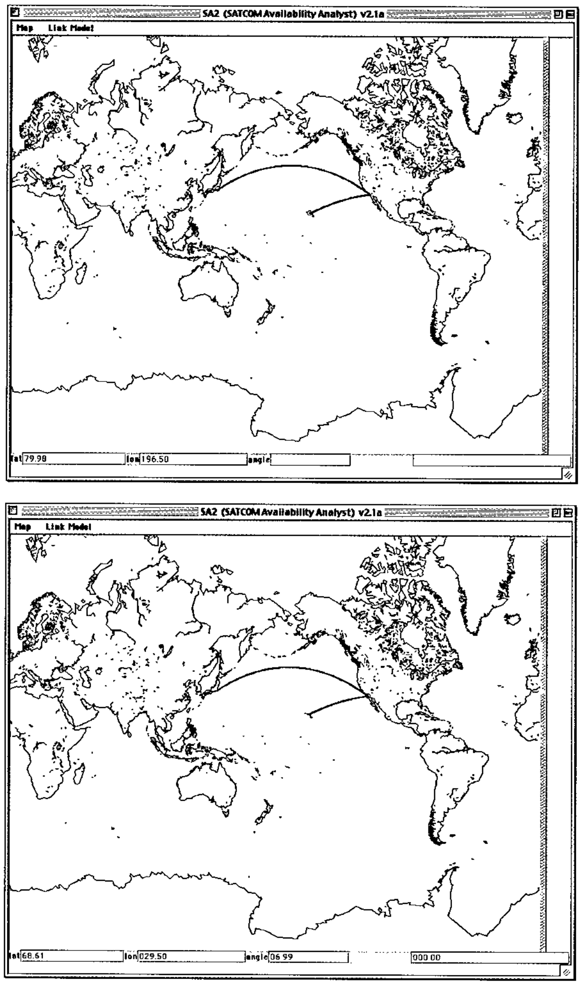
SA2 Blockage Matrix Plot
LAT 21 LON -156 HDG 0 AGF 11 GBS 0 GBS STBD

SA2 Blockage Matrix Plot
LAT 21 LON -156 HDG 0 AGF 11 GBS 0 GBS PORT or GBS STBD

FIGURE 2. SA2 SkyView display for USS *Coronado's* original GBS antenna installation locations. Top: port antenna. Middle: starboard antenna. Bottom: composite blockage for both antennas in "either antenna" mode.

The TrackView display can reflect different Sea States. It may also be animated at an accelerated speed of advance. In this animated mode, the SkyView display is slaved to the TrackView and updates as the track is traversed.

Figure 3 shows the TrackView display of the availability of the GBS due to blockage aboard USS *Coronado* (AGF 11) (corresponding to the blockage matrix in the bottom of Figure 2) in Sea State 4 along Great Circle routes between San Diego, CA, and Pearl Harbor, HI, and Yokosuka, Japan [6]. As with all SA2 TrackView displays, the blockage information plotted in Figure 3 assumes that the ship remains on Great Circle headings. With reference to the bottom of Figure 2, it is obvious that a significant number of *Coronado's* headings are blocked for GBS when the satellite in use appears above 30° elevation. This is most clearly seen using SA2's global blockage statistics and the Average Line-of-Sight Availability (ALA)View discussed in the next section.


SA2 (SATCOM Availability Analyst) v2.1a
Map
Link Model
SA2 (SATCOM Availability Analyst) v2.1a
Map
Link Model

lat 79.98 lon 196.50 angle

lat 68.61 lon 029.50 angle 06 99 000 00

FIGURE 3. SA2 TrackView display for USS *Coronado's* original GBS antenna installation locations. Red: blocked; Blue: unblocked. Satellite in use: UFO 8. Sea State 4. Top: routes outbound from San Diego. Bottom: routes inbound to San Diego.

Global Blockage Metrics

Normally, ships do not always maintain Great Circle headings while they are at sea. In general, a ship could be on any heading at a given moment depending on mission demands (e.g., flight quarters, zigzagging, etc.). Therefore, in considering the relative merits of alternative antenna installation locations, an important metric is *the percentage of headings that yield an unblocked LOS to the satellite of interest.* The following describes how SA2 calculates and displays this metric.

For any given Sea State, SA2 determines, for a combination of (1) satellite of interest, (2) ship antenna's location (or antennas' locations), and (3) ship's position and heading, whether the antenna(s) has (have) an unblocked LOS to the satellite of interest throughout the resulting ship-motion envelope. If, for a given position and heading, the satellite is visible throughout the entire ship-motion envelope, then that position is considered unblocked on that heading in the selected Sea State for the desired satellite. In performing this evaluation, by default, SA2 considers the heading blocked if the satellite is blocked at any point in the ship's motion envelope for the selected Sea State. This criterion is realistic for any bulk-encrypted link in which the encryption devices must "not miss a beat" to maintain synchronization. However, this criterion can be modified for alternative studies.

For each model cell, LOS availability is computed in the manner described in the preceding paragraph for all headings in 1° increments. The number of unblocked headings, divided by 360 and expressed as a percentage, is the ALA metric, a new figure-of-merit for analyzing blockage introduced in [6]. SA2 repeats this process for all spatial model cells. The resulting array of percentages is displayed as a color-coded map, which is SA2's ALAView. These maps show, at a glance, where a particular shipboard SATCOM terminal is unblocked at all headings, at some headings, or at no headings. Figures 4A and 4B show ALAViews for the original GBS antenna installation locations aboard *Coronado*, in calm seas, and in Sea State 6, respectively, with the GBS transponders on Ultra-High-Frequency Follow-On (UFO) satellites 8, 9, and 10 [6].

The array of ALA figures spanning the satellite constellation's field-of-regard can also be averaged, yielding a *Global ALA* (GALA) metric. GALA can be calculated in two ways: (1) the average ALA over only ocean and littoral model cells within the field-of-regard or (2) the average ALA over all model cells within field-of-regard. SA2 uses the first definition by default.

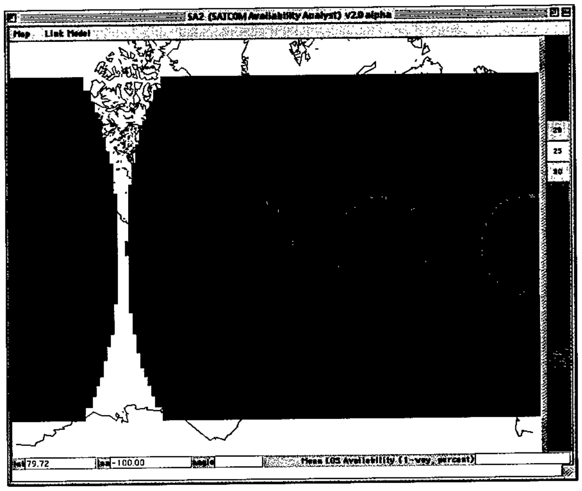

SA2 (SATCOM Availability Analyst) v2.0 alpha
Map
Link Model
20
25
30
lat 79.72
lon -100.00
angle
Mean LOS Availability (1-way, percent)

FIGURE 4A. ALAView for GBS aboard USS *Coronado*, original antenna installation locations, Sea State 0. GALA = 83.7%.

APPLICATIONS

SA2 has been used recently to analyze INMARSAT B HSD availability aboard *Ticonderoga* class (CG 47) cruisers [7], submarine HDR (SubHDR-GBS and EHF) availability aboard *Los Angeles* class (SSN 688) submarines [8], and GBS availability aboard *Coronado* [6]. A detailed account of the *Coronado* work is reported in [6]. This section summarizes some of the conclusions of the CG 47 and SSN 688 analyses.

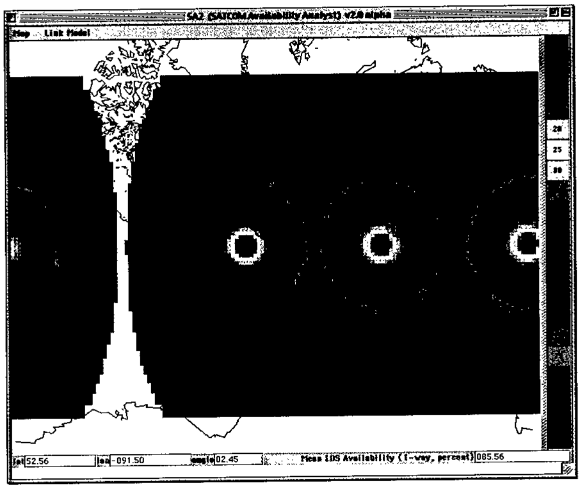
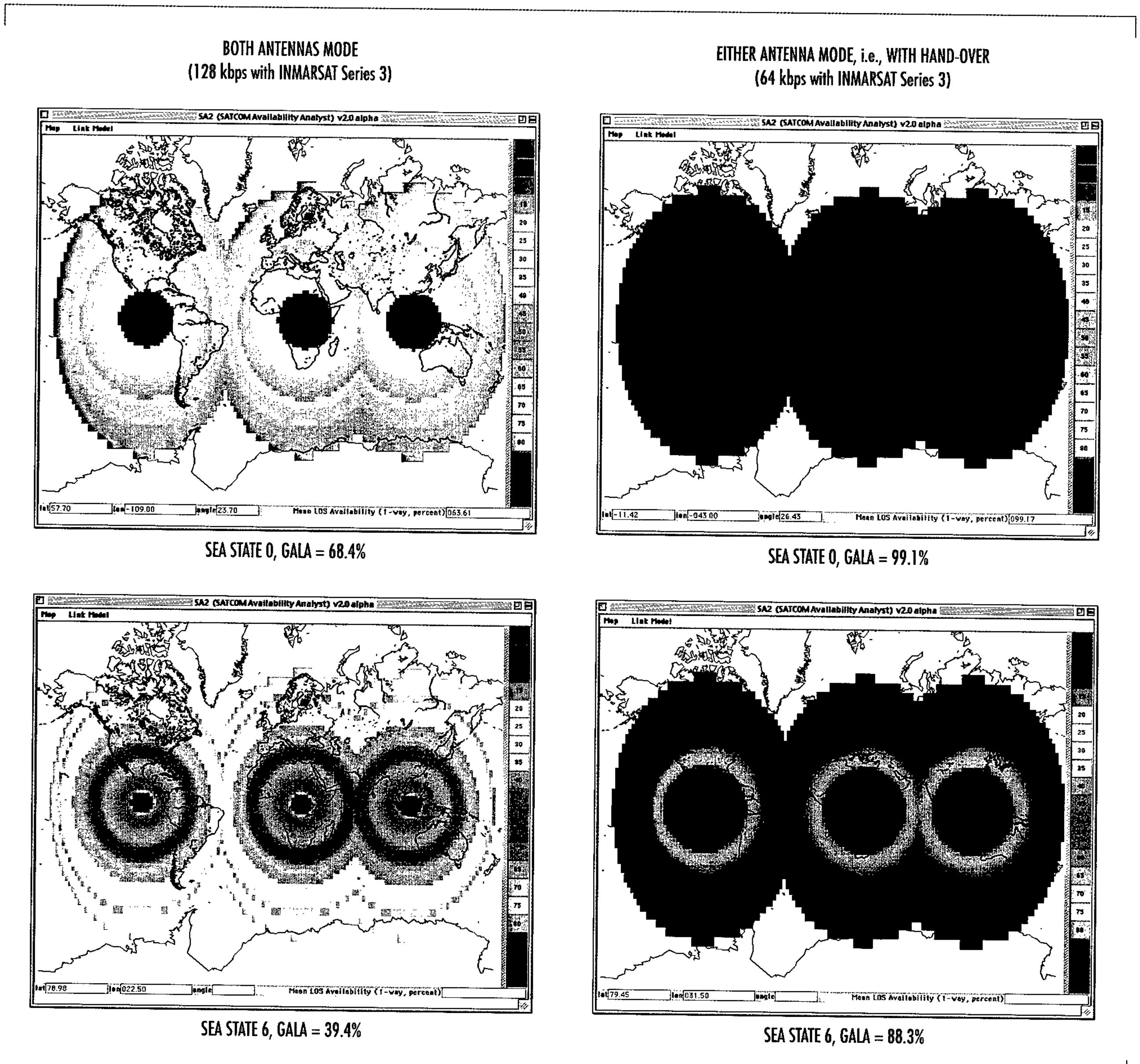

SA2 (SATCOM Availability Analyst) v2.0 alpha
Map
Link Model
20
25
30
lat 52.56
lon -091.50
angle 02.45
Mean LOS Availability (1-way, percent) 085.56

FIGURE 4B. ALAView for GBS aboard USS *Coronado*, original antenna installation locations, Sea State 6. GALA = 70.9%.

INMARSAT B HSD on the CG 47 Class

A commercial off-the-shelf (COTS) INMARSAT B HSD shipboard terminal has a single antenna and can support up to 64 kbps. To achieve 128 kbps throughput to the CG 47 and DDG 51 classes, it has been proposed to outfit each ship with two complete INMARSAT B HSD terminals. Additional INMARSAT space segment resources would be leased so that each ship would have access to an aggregate of 128 kbps by "summing" the 64 kbps channels from each terminal. Each COTS INMARSAT B HSD terminal is an independent single-antenna system. There is no tracking hand-off from one terminal's antenna to the other. Therefore, to maintain a 128-kbps aggregate, each of the two antennas must be able to view the satellite continuously as the ship maneuvers. An analysis of the impact of blockage on the availabilities that this setup would achieve was accomplished using SA2's blockage tools in "both antennas" mode (see section on Blockage Models). For comparison, "either antenna" mode was also used.

Figure 5 shows ALAViews and GALA values for Sea States 0 and 6, in INMARSAT B HSD "both antennas" mode (128 kbps) and (a hypothetical) "either antenna" mode with handover (64 kbps) using the CG 47 class INMARSAT antenna installation locations. These results clearly show



BOTH ANTENNAS MODE
(128 kbps with INMARSAT Series 3)

EITHER ANTENNA MODE, i.e., WITH HAND-OVER
(64 kbps with INMARSAT Series 3)

lat 57.70 lon -109.00 angle 23.70 Mean LOS Availability (1-way, percent) 063.61

SEA STATE 0, GALA = 68.4%

lat -11.42 lon -043 00 angle 26.43 Mean LOS Availability (1-way, percent) 099.17

SEA STATE 0, GALA = 99.1%

lat 78.98 lon 022.50 angle Mean LOS Availability (1-way, percent)

SEA STATE 6, GALA = 39.4%

lat 79.45 lon 031.50 angle Mean LOS Availability (1-way, percent)

SEA STATE 6, GALA = 88.3%

FIGURE 5. Dual-antenna INMARSAT availability on the CG 47 class analyzed with SA2's ALAView. Note: Only the three INMARSAT satellites on which leased services were offered at the time of the analysis are shown.

that a second, parallel INMARSAT B HSD terminal provides somewhat limited availability of a 128-kbps aggregate for the CG 47 class. However, the results also show that a dual-antenna INMARSAT terminal that would accomplish hand-overs between the two antennas provides an outstanding availability of 64 kbps, even in Sea State 6. After considering these results in July 2000, the Space and Naval Warfare Systems Command (SPAWAR) decided to investigate the development and acquisition of a handover-capable, dual-antenna INMARSAT B HSD terminal. Note that the next series of INMARSAT satellites, Series 4, will provide single-channel data rates up to 400 kbps, potentially making a handover-capable, dual-antenna INMARSAT terminal an even more valuable asset.

SubHDR on the SSN 688 Class

It is perhaps initially surprising that blockage is an issue for submarines, since the topside environment would appear to have no obstructions. In fact, Figure 6 shows there are several structures in close proximity to one another on the sail of the SSN 688 class. The short distances between them causes each to subtend a large solid angle as seen by the others. Furthermore, the masts and periscopes can be raised or lowered independently to variable heights.

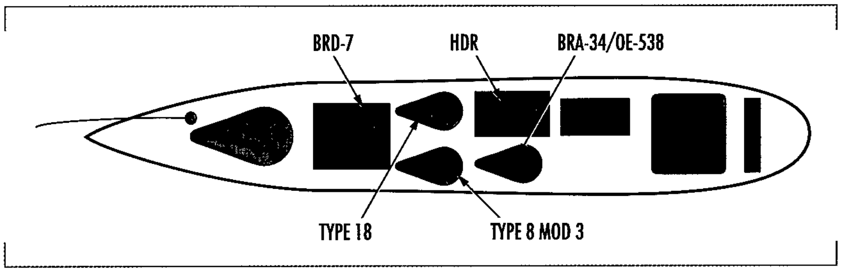


FIGURE 6. Sail configuration for USS *Providence* (adapted from [9]). The Types 8 Mod 3 and 18 are periscopes. The BRD-7 and BRA-34/OE-538 are multi-purpose masts.

The SubHDR system brings multiband SATCOM to submarines, including enhanced EHF capabilities and GBS. Early sea trials of the SubHDR mast and antenna system aboard USS *Providence* (SSN 719) revealed that from positions in the North Atlantic, the LOS to UFO 9 was sometimes blocked.

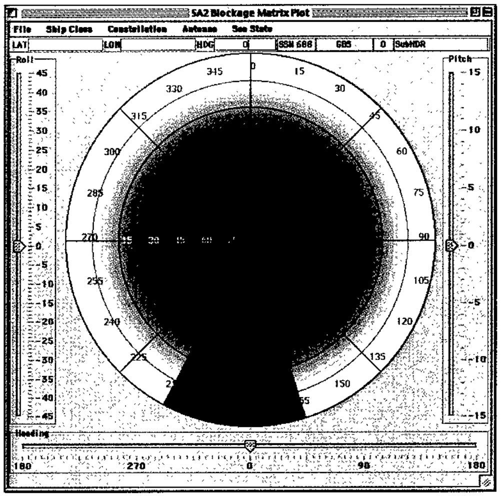


FIGURE 7. SubHDR blockage matrix when only the Type 18 periscope is raised.

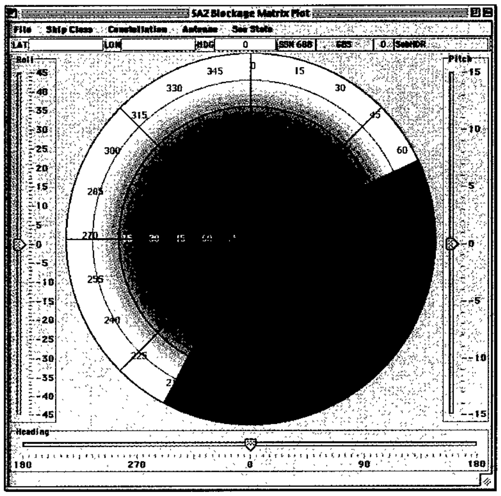


FIGURE 8. SubHDR blockage matrix when both periscopes are raised (Types 8 Mod 3 and 18) as well as both the BRA-34 and BRD-7.

For analyses of SubHDR availability, SA2 represents the periscopes and other masts independently, each as seen from the point of view of the SubHDR antenna. Thus, the number of possible blockage matrices is large, but not all of them are tactically significant. For example, by doctrine, when a submarine is at periscope depth and any mast is raised above the waterline, a periscope must also be raised. Figures 7 and 8 present examples of SubHDR blockage matrices, which correspond to the first and sixth rows of Table 1. In all cases, the SubHDR mast is lowered 14 inches from its maximum possible height to avoid blocking the periscope. Table 1 shows GALA figures for six cases of equipment raised in addition to the SubHDR mast. In Sea State 3, the SubHDR GALA figure for the GBS payloads on UFOs 8, 9, and 10, or for the EHF LDR payloads on the same spacecraft, is, at best, 90% if the Type 8 Mod 3 periscope is used and 85.7% if the Type 18 is used. Figure 9 shows an ALAView for SubHDR, assuming that only the Type 18 periscope is raised.

Clearly, it is possible to analyze blockage for submarines with SA2 by using the same tools employed for surface ships. However, at sea, blockage is a somewhat different issue for submarines than for surface ships because submariners are generally more at liberty to select blockage-free headings after reaching periscope depth (PD). For example, submariners are never concerned about orientation with respect to wind direction in order to launch or recover aircraft. Furthermore, submariners often do not stay at PD any longer than is necessary to send and receive a few queues of communications traffic. On the other hand, in rough seas at periscope depth, submariners prefer to select headings more or less directly into the swells in order to minimize roll. In any event, SA2's SkyView display is useful to submariners for selecting blockage-free headings when using SubHDR.

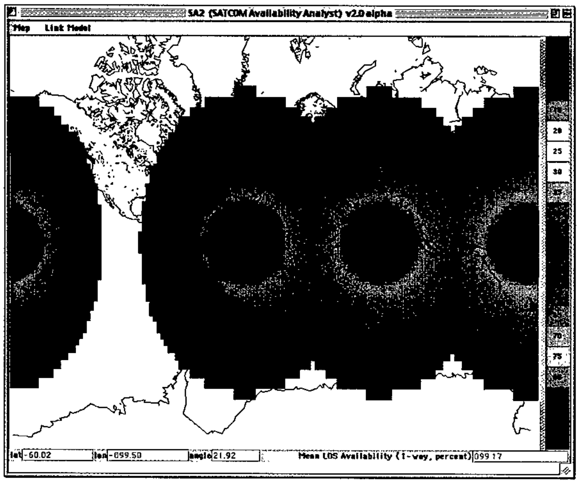
SUMMARY

SA2 combines a set of simple mathematical models of the earth and of satellite constellations, coupled with similarly straightforward models of ship motion and of superstructure blockage to produce a powerful tool for assessing the impact of blockage on shipboard and submarine SATCOM availability. All of these components were previously available in various forms, but they had never before, to our knowledge, been combined in a single, simple-to-use package.

This paper has shown that various metrics are necessary to fully describe the impact of superstructure blockage on SATCOM availability over the full set of conditions in which ships and submarines serve. We also believe that this paper and our experiences using SA2 to interact with personnel from the operational, acquisition, and RDT&E communities have demonstrated that colored graphical displays are not just desirable, but are necessary to fully convey the impact of blockage on SATCOM availability.

TABLE 1. GALA figures for GBS via SubHDR aboard the SSN 688 class for various combinations of equipment raised. The combinations represented by italicized rows, while technically possible, are not allowed by submarine doctrine.

Equipment(s) Raised (in addition to SubHDR)	GALA (%)	
	Sea State 0	Sea State 3
Type 18 Periscope	89.4	85.7
BRA-34 Mast	*83.7*	*79.5*
Type 8 Mod 3 Periscope	93.2	90.0
BRD-7 Mast	*96.4*	*94.9*
Type 18 and BRA-34	73.1	65.2
Types 8 Mod 3 and 18 plus BRA-34 and BRD-7	66.0	59.1



lat -60.02 lon -099.50 angle 21.92 Mean LOS Availability (1-way, percent) 099.17

FIGURE 9. ALAView for GBS via SubHDR aboard the SSN 688 class. Type 18 periscope is raised; SubHDR mast is lowered 14 inches from its maximum possible height. Sea State 3. GALA = 85.7%.

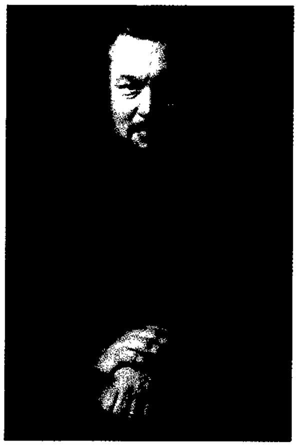
REFERENCES

1. Fitzgerald, G. and G. Bostrom. 1999. "GBS Data Mapper: Modeling Worldwide Availability of Ka-Band Links Using ITU Weather Data," *Proceedings of the IEEE Military Communications Conference (MILCOM '99)*, http://www.agreenhouse.com/society/TacCom/papers99/48_3.pdf
2. Fitzgerald, G. and G. Bostrom. 2000. "GBS Data Mapper: Modeling Worldwide Availability of Ka-Band Links Using ITU Weather Data,"*Proceedings of the 6th Ka-Band Utilization Conference*, pp. 217–224.
3. Brown, E. (RF Microsystems). 1993. "Navy EHF Program (NESP) AN/USC-38(V) Antenna Blockage Characterization," Naval Command, Control and Ocean Surveillance Center, RDT&E Division (NRaD),* San Diego, CA, (12 June), (available in the SSC San Diego EHF SATCOM In-Service Engineering Agent [ISEA] Library, catalogued as document 5000-362-19071).
4. McDonald, M. 1993. "SHF SATCOM Terminal Ship-Motion Study," TR 1578 (March), Naval Command, Control and Ocean Surveillance Center, RDT&E Division (NRaD),* San Diego, CA.
5. Department of Defense. 1986. "Interface Standard for Shipboard Systems, Ship Motion and Attitude," DoD Standard 1399, Section 301-A, (July), Washington, DC.

*now SSC San Diego

6. Axford, R. and G. Fitzgerald. 2000. "Global Broadcast Service (GBS) Blockage Assessment for USS *Coronado* (AGF 11)," TR 1842 (November), SSC San Diego, San Diego, CA.
7. Colvin, B. 2000. "INMARSAT HSD on the CG 47 and DDG 51 Classes," (31 July presentation), SSC San Diego, San Diego, CA.*
8. Fitzgerald, G. 2001. "GBS Blockage Analysis for USS *Providence*," SubHDR Test Plan Working Group (TPWG), Session 8, (23 January), Space and Naval Warfare Systems Command (SPAWAR), PMW 173 (Submarine Communications), San Diego, CA.*
9. Chief of Naval Operations. 1998. "Submarine Communications Master Plan," (April), Washington DC, p. 4-2.

❖



Roy A. Axford, Jr.

Ph.D. in Electrical Engineering, Communications Theory, and Systems, University of California at San Diego, 1995

Current Research: Technologies for wideband mobile satellite communications.

Gerald B. Fitzgerald

BA in Linguistics and Computer Science, Yale, 1977

Current Research: RF propagation modeling; imagery and SIGINT fusion; network intrusion detection.

*For further information, contact author.

Advanced Enclosed Mast/Sensor (AEM/S) System

John H. Meloling
SSC San Diego

ABSTRACT

The Advanced Enclosed Mast/Sensor (AEM/S) System is a revolutionary advancement in the topside design of Navy ships. Constructed of advanced composites, the AEM/S System is a self-supporting enclosed mast structure that provides affordable radar signature control and improved shipboard antenna system performance.

INTRODUCTION

The Advanced Enclosed Mast/Sensor (AEM/S) System uses advanced composites to produce a mast structure that encloses the existing legacy antenna systems of the ship. This enclosure consists of a composite sandwich structure that supports all internal decks, antennas, and ballistic cable trunks. Embedded within the composite sandwich are frequency selective surface (FSS) layers that filter electromagnetic waves. This filtering allows transmission and reception at desired frequencies while rejecting threat radar signals. Once these electromagnetic characteristics are designed into the composite sandwich, the mast structure can be shaped to reduce the radar cross section (RCS).

The AEM/S technology has many advantages. AEM/S provides affordable signature control of legacy antenna systems. Developing and fielding new antenna systems is a long and costly process. The AEM/S System provides a near-term means of reducing the RCS of ships. Many of the newer antenna systems under development plan to use phased array antennas. The faceted nature of the AEM/S structure provides the necessary flat surfaces for mounting these future systems. The performance of the enclosed shipboard antennas is improved over conventional metallic masts because there is less blockage of the antenna. Maintenance of the enclosed antennas is reduced because the antennas are not exposed to adverse weather, wind loading, salt water, or stack gases. Less maintenance directly reduces costs over the entire service life of the ship.

AEM/S SYSTEM ATD

The AEM/S concept was demonstrated through the Office of Naval Research (ONR) AEM/S System Advanced Technology Demonstration (ATD) project. This FY 1995 ATD demonstrated the ability to design and fabricate enclosed mast structures for Navy ships. Figure 1 shows the

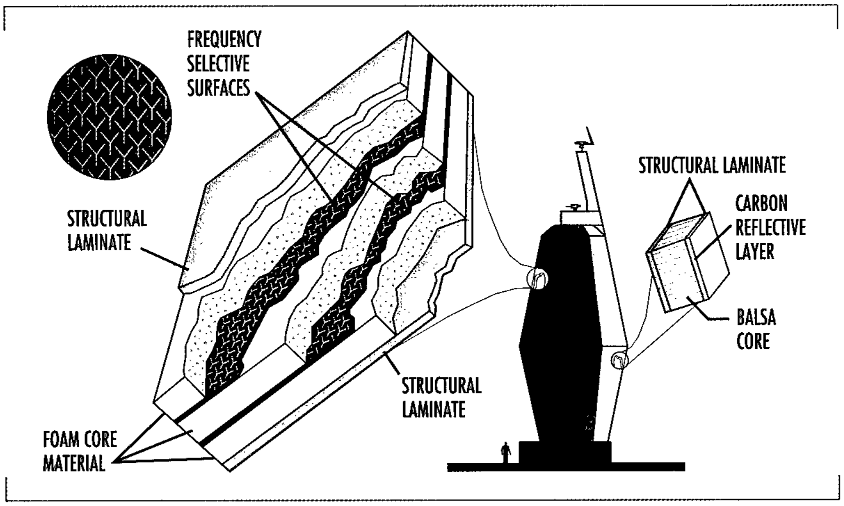


FIGURE 1. AEM/S System ATD sandwich construction concept.

AEM/S ATD configuration, with the FSS structure on the top and a balsa core reflective composite on the bottom. This ATD fused advances in electromagnetics, signature reduction, structures, materials, and manufacturing technologies. The all-composite, self-supporting enclosure is approximately 100 feet tall, 36 feet in diameter, and 40 tons in weight.

SSC San Diego played a major role in the development and the success of the AEM/S System by performing all of the electromagnetic design and development for the program. SSC San Diego's involvement included designing and validating the FSS radomes, handling antenna integration issues such as antenna placement and electromagnetic compatibility, developing new antenna designs such as the Integrated High-Frequency Antenna, and performing antenna performance predictions and measurements of major enclosed radar systems.

The FSS radome design process required artful compromise between electromagnetic, mechanical, and material engineering disciplines. Optimum mast wall design was achieved through tradeoffs between enclosed antenna system performance in the passband and the threat signal rejection level in the stop band. Also, mechanical consideration of strength bound the acceptable ranges of the composite skin and core thickness. Materials were selected for their electrical properties, mechanical strength, thermal properties, and cost.

Electromagnetic compatibility is designed into the mast through proper antenna placement. This compatibility is achieved by using the conducting decks as shielding, using the filtering characteristics of the radomes, and designing the structure to minimize electromagnetic interference while maximizing coverage.

SSC San Diego conceived the idea of mounting a high-frequency (HF) antenna to the inside surface of a radome during the research phase prior to the start of the AEM/S ATD. Eventually called the Integrated High-Frequency Antenna (IHFA), the concept offered a novel approach to the design of HF antennas for Navy ships in that (1) the radome structure provides the necessary height and volume to produce a good HF antenna, and (2) by mounting the antenna to the inside surface of the radome, the antenna cannot be seen by threat radars.

SSC San Diego also developed a new capability for antenna performance evaluation on Navy ships during the AEM/S ATD. With assistance from The Ohio State University ElectroScience Laboratory, new computer modeling tools were developed for the analysis of radome-enclosed antennas. This capability has been validated using scale-model and full-scale antenna pattern measurements.

In 1997, the AEM/S ATD culminated with the installation and at-sea testing of the mast on USS *Arthur W. Radford* (DD 968). Figure 2 shows *Arthur W. Radford* at sea with the AEM/S mast installed. This mast provides superior antenna system performance. The AEM/S mast also provides significant RCS reduction, reduced antenna system maintenance, and reduced life-cycle costs.

AEM/S FOR LPD 17

While the ATD mast was being fabricated and installed, members of the ATD project team and the Naval Sea Systems Command LPD 17 program office began discussion of potential advantages of AEM/S. Because the ATD technology showed performance and maintenance advantages for the LPD 17 platform,* a risk mitigation project to address technology transition and design issues for LPD 17 was initiated.

As with the AEM/S System ATD, SSC San Diego has played a major role in the development and the success of the AEM/S for the LPD 17 program. SSC San Diego performed all of the electromagnetic design and development for the program. Involvement has included designing and validating the FSS radomes, handling antenna integration issues such as antenna placement and electromagnetic compatibility, developing the IHFA designs, and performing antenna performance predictions and measurements of major enclosed systems.

The design of the AEM/S for LPD 17 involves the design of two separate masts. As such, each mast has different requirements, different antenna systems, and, therefore, different challenges. In both cases, the performance requirements were more difficult to meet than those of the ATD radome, largely because of the considerable increase in the signature requirements of the masts. Another challenge for the radome for the aft mast is meeting the more stringent requirements of the SPS-48E radar. SPS-E is a higher frequency, higher gain radar that is sensitive to any variations caused by the surrounding ship structure. These requirements, in addition to the extreme structural requirements imposed by the height of the enclosure (approximately 12 meters), suggest the extraordinary interdisciplinary cooperation necessary to obtain an optimum design.

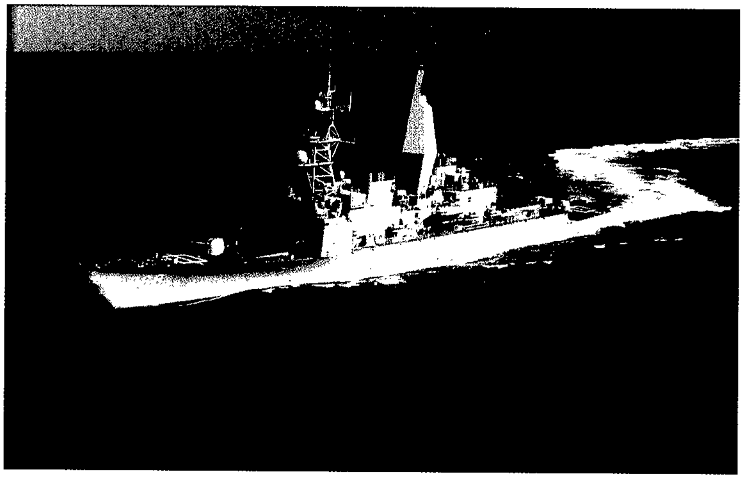


FIGURE 2. AEM/S ATD at sea on USS *Arthur W. Radford* (DD 968).

Figure 3 shows an artist's conception of the LPD 17 with the AEM/S masts installed.

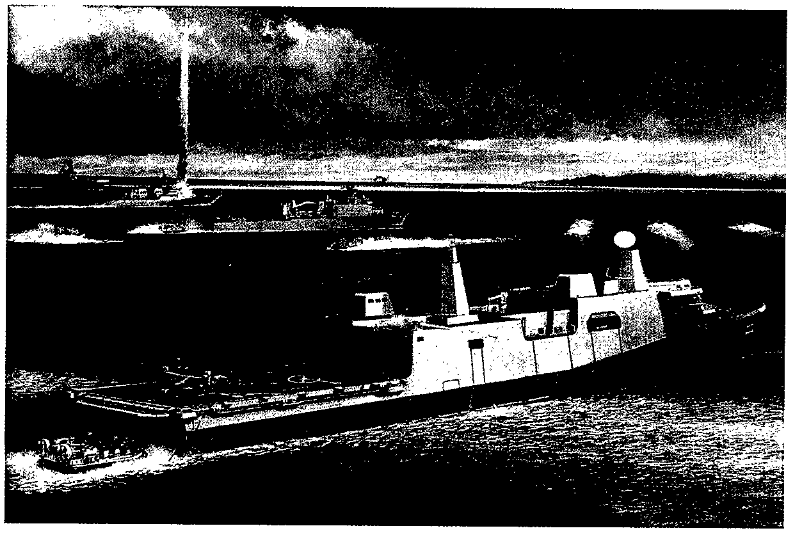


FIGURE 3. Artist's conception of the LPD 17 with AEM/S masts.

Another of the many challenges associated with the LPD 17 AEM/S is the design of IHFAs for low- and high-band transmission. Under the ATD, only the design of a high-band antenna was treated. The low-band IHFA requires a radome structure that provides the necessary height and volume to produce a good HF antenna. These challenges were

*Landing Platform Dock 17 (LPD 17), *San Antonio* class, is the latest class of amphibious force ship for the U.S. Navy. The first ship, USS *San Antonio* (LPD 17) is currently under construction.

met by the inclusion of the high-band IHFA in the shorter forward mast, and the low-band antenna in the taller aft mast.

Predicting antenna performance for enclosed antennas was particularly difficult for the SPS-48E radar. SPS-48E is a volume search radar with very high gain and low sidelobes. Predicting performance has proven to be one of the most challenging aspects of the LPD 17 design. However, a good understanding of the performance of the enclosed antenna has been obtained through advanced computer modeling and component-level measurements.

In 1999, the AEM/S risk mitigation effort culminated with the official change of the design from the contract metal masts to the AEM/S masts; this was a milestone comparable in significance to the installation of the original ATD mast on *Arthur W. Radford.* Since that time, work has continued in all areas to obtain designs that are ready to meet the production schedule of the lead ship.

CONCLUSION

The AEM/S System is a unique U.S. Navy program that has encompassed research and development, an Advanced Technology Demonstration (ATD), and new ship construction (LPD 17). This successful transition of technology has made the AEM/S System program one of the most successful programs of the last decade. The highly integrated and consensus-managed team of Navy and industry experts has made this program successful. The program's success and numerous benefits will encourage the Navy to continue implementing the AEM/S System and its associated technologies.

❖

John H. Meloling

Ph.D. in Electrical Engineering, Ohio State University, 1994

Current Research: Frequency-selective surfaces; radomes; absorbers; high-frequency electromagnetics.

Seaweb Underwater Acoustic Nets

Joseph A. Rice, Robert K. Creber,
Christopher L. Fletcher, Paul A. Baxley,
Kenneth E. Rogers, and Donald C. Davison
SSC San Diego

ABSTRACT

Seaweb networks use digital signal processor (DSP)-based telesonar underwater acoustic modems to interconnect fixed and mobile nodes. Backbone nodes are autonomous, stationary sensors and telesonar repeaters. Peripheral nodes include unmanned undersea vehicles (UUVs) and specialized devices such as low-frequency sonar projectors. Gateway nodes provide interfaces with command centers afloat, submerged, ashore, and aloft, including access to terrestrial, airborne, and space-based networks. Seaweb command, control, communications, and navigation (C^{3}N) technology coordinates deployable assets for accomplishing given missions in littoral ocean environments. A series of annual experiments drives seaweb technology development by implementing increasingly sophisticated wide-area networks of deployable autonomous undersea sensors.

INTRODUCTION

Digital signal processor (DSP) electronics and the application of digital communications theory have substantially advanced the underwater acoustic telemetry state of the art [1]. A milestone was the introduction of a DSP-based modem [2] sold as the Datasonics ATM850 [3 and 4] and later identified as the first-generation telesonar modem. To promote further development of commercial off-the-shelf (COTS) telesonar modems, the U.S. Navy invested small business innovative research (SBIR) funding and Navy laboratory support with expectations that energy-efficient, inexpensive telesonar modems would spawn autonomous undersea systems [5]. Steady progress resulted in the second-generation telesonar modem [6], marketed as the Datasonics ATM875. Encouraged by the potential demonstrated with the ATM875, the Navy funded the advanced development of a third-generation telesonar modem [7] designated the Benthos ATM885.

Seaweb is an organized network for command, control, communications, and navigation (C^{3}N) of deployable autonomous undersea systems. Seaweb functionality implemented on telesonar hardware shows enormous promise for numerous ocean applications.

Offboard seaweb nodes of various types may be readily deployed from high-value platforms including submarine, ship, and aircraft, or from unmanned undersea vehicles (UUVs) and unmanned aerial vehicles (UAVs). The architectural flexibility afforded by seaweb wireless connections permits the mission planner to allocate an arbitrary mix of node types with a node density and area coverage appropriate for the given telesonar propagation conditions and for the mission at hand.

The initial motivation for seaweb is a requirement for wide-area undersea surveillance in littoral waters by means of a deployable autonomous distributed system (DADS) such as that shown in Figure 1. Future sensor nodes in a DADS network generate concise antisubmarine warfare (ASW) contact reports that seaweb will route to a master node for field-level data fusion [8]. The master node communicates with manned command centers via gateway nodes such as a sea-surface buoy radio-linked with space satellite networks, or a ship's sonar interfaced to an onboard seaweb server.

DADS operates in 50- to 300-m waters with node spacing of 2 to 5 km. Primary network packets are contact reports with about 1000 information bits [9]. DADS sensor nodes asynchronously produce these packets at a variable rate dependent on the receiver operating characteristics for a particular sensor suite and mission. Following *ad hoc* deployments, DADS relies on the seaweb network for self-organization including node identification, clock synchronization on the order of 0.1 to 1.0 s, node geo-localization on the order of 100 m, assimilation of new nodes, and self-healing following node failures. Desired network endurance is up to 90 days.

DADS is a fixed grid of inexpensive interoperable nodes. This underlying cellular network architecture is well suited for supporting an autonomous oceanographic sampling network (AOSN) [10], including C^{3}N for autonomous operations with UUV mobile nodes.

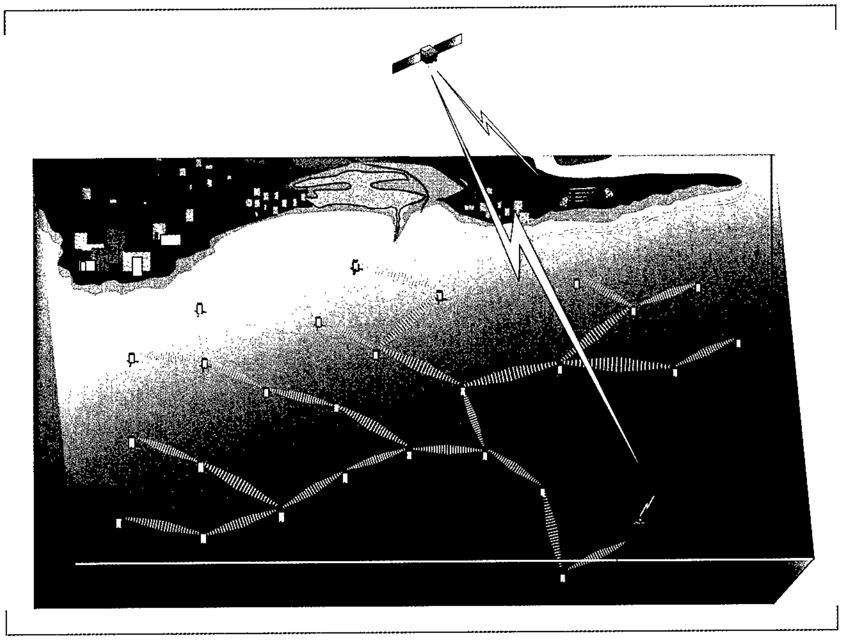


FIGURE 1. Seaweb underwater acoustic networking enables C^{3}N for DADS and other deployable autonomous undersea systems. Gateways to manned control centers include radio links to space or shore and telesonar links to ships.

CONCEPT OF OPERATIONS

Telesonar wireless acoustic links interconnect distributed undersea instruments, potentially integrating them as a unified resource and extending "net-centric" operations into the undersea environment.

Seaweb is the realization of such an undersea wireless network [11] of fixed and mobile nodes, including various interfaces to manned command centers. It provides the C^{3}N infrastructure for coordinating appropriate assets to accomplish a given mission in an arbitrary ocean environment.

The seaweb backbone is a set of autonomous, stationary nodes (e.g., deployable surveillance sensors, repeaters). Seaweb peripherals are mobile nodes (e.g., UUVs, including swimmers, gliders, and crawlers) and specialized nodes (e.g., bistatic sonar projectors).

Seaweb gateways provide connections to command centers submerged, afloat, aloft, and ashore. Telesonar-equipped gateway nodes interface seaweb to terrestrial, airborne, and space-based networks. For example, a telesonobuoy serves as a radio/acoustic communications (racom) interface, permitting satellites and maritime patrol aircraft to access submerged, autonomous systems. Similarly, submarines can access seaweb with telesonar signaling through the WQC-2 underwater telephone band or other high-frequency sonars [12]. Seaweb provides the submarine commander with digital connectivity at speed and depth and with bidirectional access to all seaweb-linked resources and distant gateways.

A seaweb server resides at manned command centers and is the graphical user interface to the undersea network as shown in Figure 2. The server archives all incoming data packets and provides read-only access to client stations via the Internet. A single designated "super" server controls and reconfigures the network.

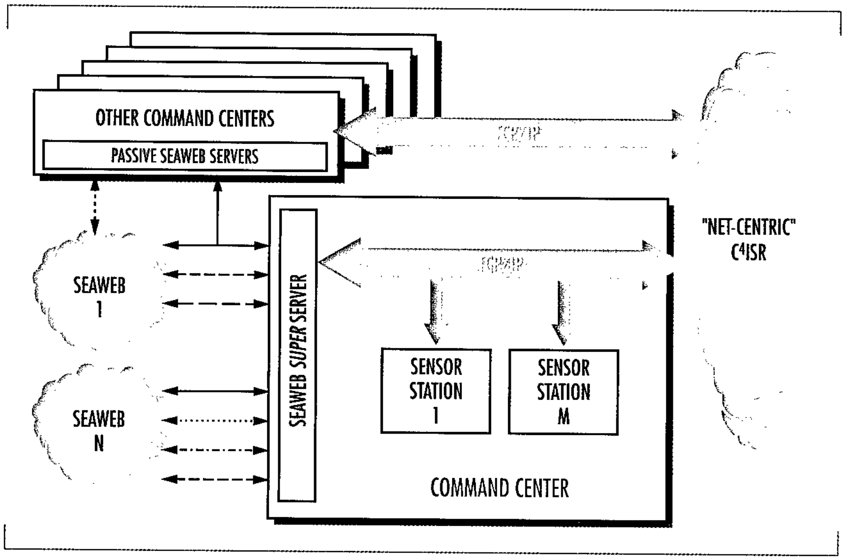


FIGURE 2. Seaweb extends modern "net-centric" interconnectivity to the undersea realm. Wireless underwater networks include gateway nodes with radio, acoustic, wire, or fiber links to manned command centers where a seaweb server provides a graphical user interface. Command centers may be aboard ship, submarine, aircraft, or ashore. They may be geographically distant and connected to the gateway node via space satellite or terrestrial Internet. At the designated command center, a seaweb "super" server manages and controls the undersea network. All seaweb servers archive seaweb packets and provide data access to client stations via the Internet. A single designated super server controls and reconfigures the network.

Low-bandwidth, half-duplex, high-latency telesonar links limit seaweb quality of service. Occasional outages from poor propagation or elevated noise levels can disrupt telesonar links [13]. Ultimately, the available energy supply dictates service life, and battery-limited nodes must be energy conserving [14]. Moreover, seaweb must ensure transmission security by operating with low bit-energy per noise-spectral-density (E_b/N_0) and by otherwise limiting interception by unauthorized receivers. Seaweb must therefore be a revolutionary information system bound by these constraints.

Simplicity, efficiency, reliability, and security are the governing design principles. Half-duplex handshaking [15] asynchronously establishes adaptive telesonar links [16] as shown in Figure 3. The initiating node transmits a request-to-send (RTS) waveform with a frequency-hopped, spread-spectrum (FHSS) [17] pattern or direct-sequence spread-spectrum (DSSS) [18] pseudo-random carrier uniquely addressing the intended receiver. (Alternatively, the initiating node may transmit a universal code for broadcasting or when establishing links with unknown nodes.) The addressed node detects the request and awakens from an energy-conserving sleep state to demodulate. Further processing of the RTS signal provides an estimate of the channel scattering function and signal excess. The addressed node then acknowledges receipt with a FHSS or DSSS acoustic response. This clear-to-send (CTS) reply specifies appropriate modulation parameters for the ensuing message packets based upon the measured channel conditions. Following this RTS/CTS handshake, the initiating node transmits the data packet(s) with nearly optimal bit-rate, modulation, coding, and source level.

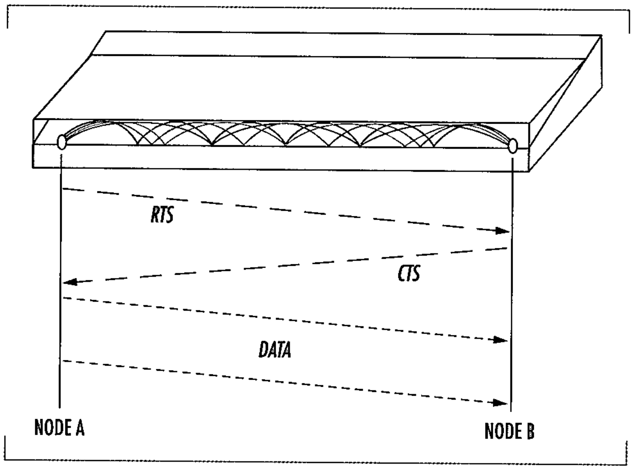


FIGURE 3. Telesonar handshake protocol for data transfer involves Node A initiating a request-to-send (RTS) modulated with a channel-tolerant, spread-spectrum pattern uniquely associated with intended receiver, Node B. So addressed, Node B awakens and demodulates the fixed-length RTS packet. Node B estimates the channel parameters using the RTS as a probe signal. Node B responds to A with a fixed-length clear-to-send (CTS) that fully specifies the modulation parameters for the data transfer. Node A then sends the data packet(s) with optimal source level, bit-rate, modulation, and coding. If Node B receives corrupted data, it initiates a selective automatic repeat request (ARQ).

Spread-spectrum modulation is consistent with the desire for asynchronous multiple access to the physical channel using code-division multiple-access (CDMA) networking [19]. Nevertheless, the seaweb concept does not exclude time-division multiple-access (TDMA) or frequency-division multiple-access (FDMA) methods, and is pursuing hybrid schemes suited to the physical-layer constraints. In a data transfer, for example, the RTS/CTS exchange might occur as an asynchronous CDMA dialog in which the data packets are queued for transmission during a time slot or within a frequency band such that collisions are avoided altogether.

The seaweb architecture of interest includes the physical layer, the media-access-control (MAC) layer, and the network layer. These most fundamental layers of communication functionality support higher layers that will tend to be application-specific.

At the physical layer, an understanding of the transmission channel is obtained through at-sea measurements and numerical propagation models. Knowledge of the fundamental constraints on telesonar signaling translates into increasingly sophisticated modems. DSP-based modulators and demodulators permit the application of modern digital communication techniques to exploit the unique aspects of the underwater channel. Directional transducers further enhance the performance of these devices [20].

The MAC layer supports secure, low-power, point-to-point connectivity, and the telesonar handshake protocol is uniquely suited to wireless half-duplex networking with slowly propagating channels. Handshaking permits addressing, ranging, channel estimation, adaptive modulation, and power control. The seaweb philosophy mandates that telesonar links be environmentally adaptive [21], with provision for bidirectional asymmetry.

Network supervisory algorithms can execute either at an autonomous master node or at the seaweb server. Seaweb provides for graceful failure of network nodes, addition of new nodes, and assimilation of mobile nodes. Essential by-products of the telesonar link are range measurement, range-rate measurement, and clock-synchronization. Collectively, these C^3N features support network initialization, node localization, route configuration, resource optimization, and maintenance.

DEVELOPMENTAL APPROACH

Given the DADS performance requirements, seaweb research is advancing telesonar modem technology for reliable underwater signaling by addressing the issues of (1) adverse transmission channel; (2) asynchronous networking; (3) battery-energy efficiency; (4) transmission security; and (5) cost.

Despite an architectural philosophy emphasizing simplicity, seaweb is a complex system and its development is a grand challenge. The high cost of sea testing and the need for many prototype nodes motivate extensive engineering system analysis following the ideas of the previous section.

Simulations using an optimized network engineering tool (OPNET) with simplified ocean acoustic propagation assumptions permit laboratory refinement of networking protocols [22] and initialization methods [23]. Meanwhile, controlled experimentation in actual ocean conditions incrementally advances telesonar signaling technology [24].

Seaweb experiments implement the results from these research activities with a periodic concentration of resources in prolonged ocean experiments.

The annual seaweb experiments validate system analysis and purposefully evolve critical technology areas such that the state-of-the-art advances with greater reliability, functionality, and quality of service. The objective of the seaweb experiments is to exercise telesonar modems in networked configurations where various modulation and networking algorithms can be assessed. In the long-term, the goal is to provide for a self-configuring network of distributed assets, with network links automatically adapting to the prevailing environment through selection of the optimum transmit parameters.

A full year of hardware improvements and in-air network testing helps ensure that the incremental developments tested at sea will provide tractable progress and mitigate overall developmental risk. In particular, DADS relies on the annual seaweb engineering experiments to push telesonar technology for undersea wireless networking. After the annual seaweb experiment yields a stable level of functionality, the firmware product can be further exercised, and refinements can be instituted during DADS system testing and by spin-off applications throughout the year. For example, in year 2001, seaweb technology enables the March–June *FRONT-3* ocean observatory on the continental shelf east of Long Island, NY [25]. These applications afford valuable long-term performance data that are not obtainable during seaweb experiments when algorithms are in flux and deployed modems are receiving frequent firmware upgrades.

The *Seaweb '98*, *'99*, and *2000* operating area in Buzzards Bay is framed in Figure 4. An expanse of 5- to 15-m shallow water is available for large-area network coverage with convenient line-of-sight radio access to Datasonics and Benthos facilities in western Cape Cod, MA. A shipping channel extending from the Bourne Canal provides episodes of high shipping noise useful for stressing the link signal-to-noise ratio (SNR) margins. The seafloor is patchy with regions of sand, gravel, boulders, and exposed granite.

Figure 5 shows *Seaweb '98*, *'99*, and *2000* modem rigging. Experiments occur during August and September when weather is conducive to regular servicing of deployed network nodes.

A representative sound-speed profile inferred from a conductivity-temperature-depth (CTD) probe during *Seaweb '98* is shown in Figure 6. For observed August and September sound-speed profiles, ray tracing suggests maximum direct-path propagation to ranges less than 1000 m, as Figure 7 shows. Beyond this distance, received acoustic energy is via boundary forward scattering. Ray tracing further indicates that received signal energy at significant ranges is attributable

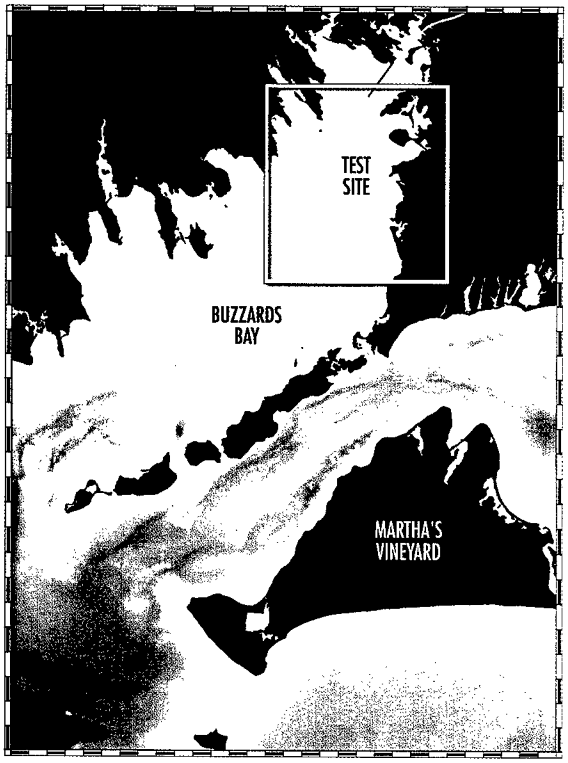


FIGURE 4. The test site for *Seaweb '98*, *'99*, and *2000* is northern Buzzards Bay, MA. Water depth is 5 to 15 m.

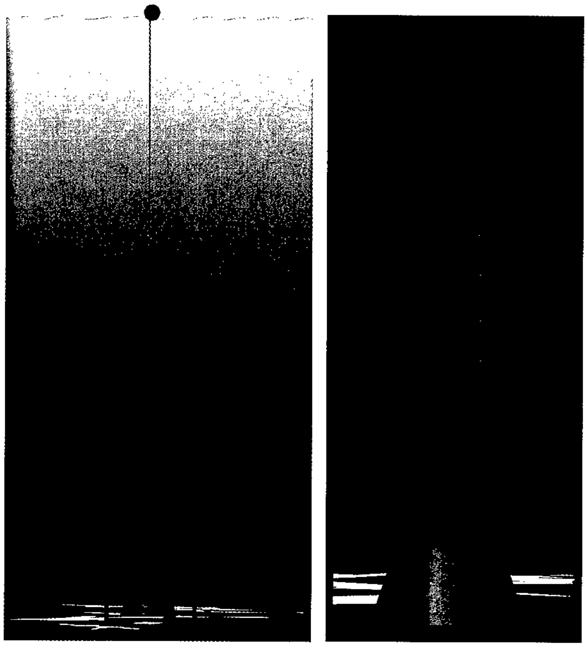


FIGURE 5. *Seaweb '98*, *'99*, and *2000* modems are deployed in Buzzards Bay with concrete weight, riser line, and surface float. The shallow water and simple rigging permit a small craft to rapidly service the network. Servicing includes battery replacement and modem firmware downloads.

to a very small near-horizontal continuum of projector elevation launch angles. Figure 8 presents predicted impulse responses for 10 ranges, each revealing multipath spreads of about 10 ms [26]. All ranges are considered "long" with respect to water depth. Summer afternoon winds and boat traffic regularly roughen the sea surface, increasing scattering loss and elevating noise levels.

SEAWEB '98 EXPERIMENT

Seaweb '98 led off a series of annual ocean experiments intended to progressively advance the state of the art for asynchronous, non-centralized networking. *Seaweb '98* used the Datasonics ATM875 second-generation telesonar modem [27] recently available as the product of a Navy SBIR Phase-2 contract.

The ATM875 normally uses 5 kHz of acoustic bandwidth with 120 discrete multiple-frequency shift keying (MFSK) bins configured to carry six Hadamard codewords of 20 tones each. Interleaving the codewords across the band increases immunity to frequency-selective fading, and Hadamard coding yields a frequency diversity factor of 5 for adverse channels having low or modest spectral coherence. This standard ATM875 modulation naturally supports three interleaved FDMA sets of 40 MFSK tonals and two codewords each. To further reduce multi-access interference (MAI) between sets, half the available bandwidth capacity provided additional guardbands during *Seaweb '98*. Thus, only 20 MFSK tonals composing one Hadamard codeword formed each FDMA set. The *Seaweb '98* installation was three geographic clusters of nodes with FDMA sets "A" through "C" mapped by cluster. For example, all nodes in cluster A were assigned the same FDMA carrier set for reception. Each cluster contained a commercial oceanographic sensor at a leaf node asynchronously introducing data packets into the network. This FDMA architecture was an effective multi-access strategy permitting simultaneous network activity in all three clusters without MAI [28]. A drawback of FDMA signaling is the inefficient use of available bandwidth. *Seaweb '98* testing was based on a conservative 300-bit/s modulation to yield a net FDMA bit-rate of just 50 bit/s. This was an acceptable rate since the *Seaweb '98* objective was to explore networking concepts without excessive attention to signaling issues. Within a cluster, TDMA was the general rule broken only by deliberate intrusion from the command center.

The gateway node is an experimental Navy racom buoy (Figure 9). The "master" node was installed approximately 1500 m from the gateway node. Gateway and master nodes formed cluster C, and so received and demodulated only the FDMA carriers of set C. Exercising the link between gateway and master nodes during various multi-hour and multi-day

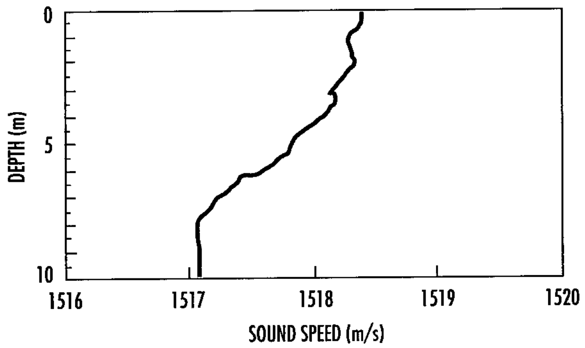


FIGURE 6. Sound-speed profiles calculated from conductivity and temperature probes are generally downward refracting during August–September at the *Seaweb'98*, *'99*, and *2000* site. This sound-speed profile, 1 of 14 obtained during *Seaweb '98*, is typical. The sound-speed gradient evident here is caused by summertime sea-surface warming.

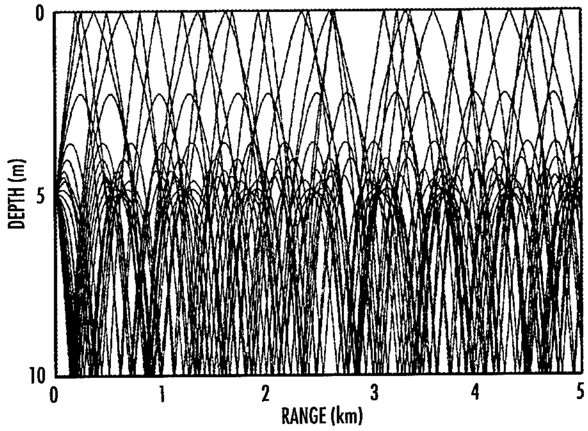


FIGURE 7. *Seaweb '98* propagation refracts downward in response to the vertical sound-speed gradient observed in Figure 6. Rays traced from a ±2.5° vertical fan of launch elevation angles model the telesonar sound channel for transmitter at 5-m depth. A parametric modeling study assessing the dependence of modem depth for this environment confirmed the general rule that long-range signaling in downward-refracting, non-ducted waters is optimized with modems placed nearer the seafloor. Hence, *Seaweb '98*, *'99*, and *2000* modem transducers are generally about 2 m above the bottom.

periods yielded link statistics for improving the wake-up and synchronization schemes in the modem receiver acquisition stage. This point-to-point testing identified specific suspected problems in the fledgling ATM875 implementation, and firmware modifications improved the success of packet acquisition from 80% to 97%.

Installation of a three-node subset of cluster A added a relay branch around Scraggy Neck, a peninsula protruding into Buzzards Bay. An Ocean Sensors CTD produced data packets relayed via each of the intervening A nodes to the master node, and then on to the gateway node. Each relay link was about 1500 m in range. Direct addressing of cluster-A nodes from the gateway node confirmed the existence of reliable links to all but the outermost node. Remarkably, a reliable link existed between two nodes separated by 3.6 km in spite of shoaling to 1 to 2 m in intervening waters! Various network interference situations were intentionally and unintentionally staged and tested until this simple but unprecedented relay geometry was well understood.

These early tests realized an unexpected benefit of the gateway link between the racom buoy and the radio-equipped workboat. End-to-end functionality of a newly installed node could be immediately verified. Field personnel would use a deck unit and the gateway node to test the network circuit that included the new modem as an intermediate node, or they would bidirectionally address the new modem via just the gateway route. Effectively, the workboat was a mobile node in the network equipped with both telesonar and gateway connections.

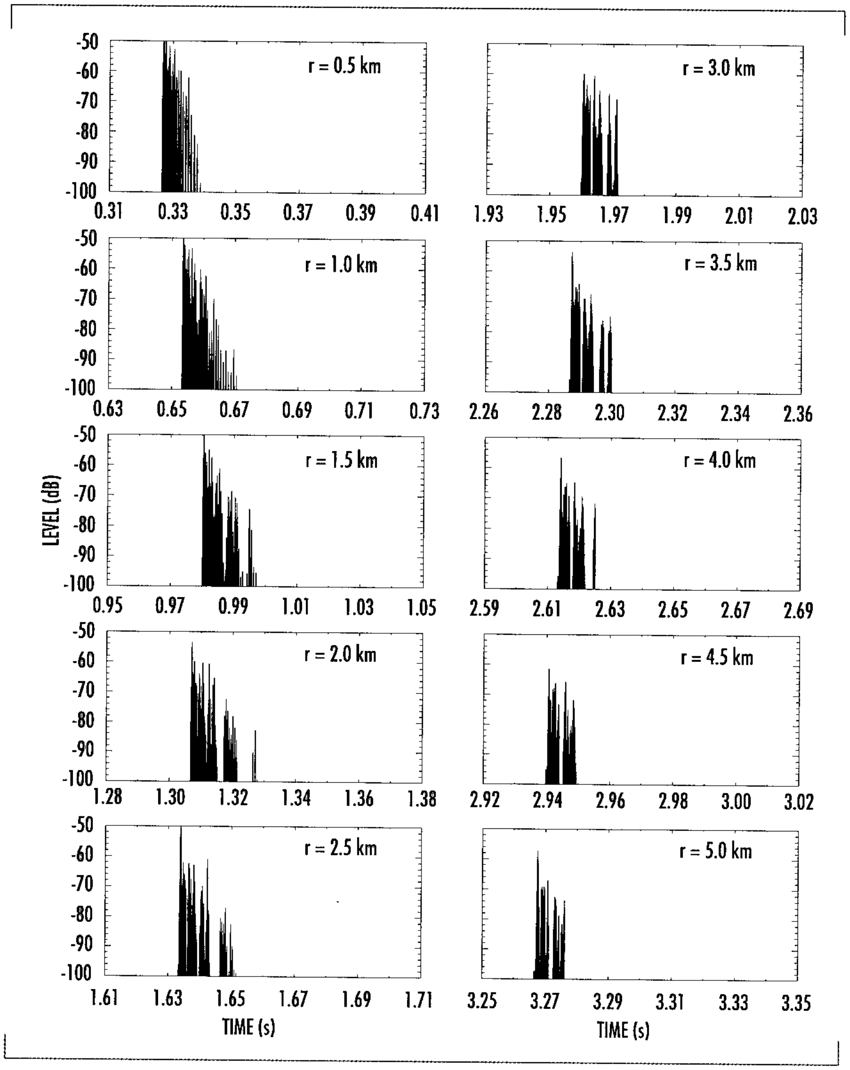


FIGURE 8. For a 10-m deep *Seaweb '99* channel, a 2-D Gaussian beam model predicts impulse responses for receivers located at 10 ranges, r. Response levels are in decibels referenced to a 0-dB source. Multipath spread is about 10 ms. Note the *Seaweb '98*, *'99*, and *2000* working ranges are hundreds of times greater than the water depths, and boundary interactions are therefore complex. For rough sea floor and sea surface, the 2-D model approximation must give way to 3-D forward scattering, and the predicted response structures will instead be smeared by out-of-plane propagation. *Seaweb 2000* testing includes channel probes designed to directly measure channel scattering functions with receptions recorded at various ranges by telesonar test beds. These channel measurements support analysis of experimental signaling and help calibrate an experimental 3-D Gaussian beam model under development for telesonar shallow-water performance prediction.

At this point, associates from the National Oceanic and Atmospheric Agency (NOAA) and Naval Surface Warfare Center (NSWC) visited *Seaweb '98*. A boat delivered them far into Buzzards Bay, where a hydrophone (deployed over the side with a telesonar deck unit) turned the boat into just such a mobile network node. The visitors typed messages, which were relayed through the network and answered by personnel at the ashore command center.

Next, a branch was added to cluster A with a Falmouth Scientific 3-D current meter and CTD. Network contention was studied by having the two cluster-A sensor nodes generate packets at different periods such that network collisions would occur at regular intervals with intervening periods of non-colliding network activity.

Finally, cluster B was introduced to the network with internode separations of 2 km. A third device generated data packets. With all available network nodes installed and functioning, the remaining few days involved a combination of gradually arranging network nodes with greater spacing as charted in Figure 10, and of doing specialized signal testing with the telesonar test bed [29]. In addition, the telesonar test bed was deployed in the center of the network for five data-acquisition missions and recorded 26 hours of acoustic network activity. The test bed also included a modem, permitting it to act as the tenth network node and giving ashore operators the ability to remotely control and monitor test-bed operations. The test-bed node provides raw acoustic data for correlation with automatic modem diagnostics, providing opportunity to study failure modes using recorded time series.

Seaweb '98 demonstrated the feasibility of low-cost distributed networks for wide-area coverage. During 3 weeks of testing in September, the network performed reliably through a variety of weather and noise events. Individual network links spanned horizontal ranges hundreds of water depths in length. The *Seaweb '98* network connected widely spaced autonomous modems in a binary-tree topology with a master node at the base and various oceanographic instruments at outlying leaf nodes. Also connected to the master node was an acoustic link to a gateway buoy, providing a line-of-sight digital radio link to the command center ashore. The network transported data packets acquired by the oceanographic instruments through the network to the master node, on to the gateway node, and then to the command center. The oceanographic instruments and modems generally operated according to preprogrammed schedules designed to periodically produce network collisions, and personnel at the command center or aboard ship also remotely controlled network nodes in an asynchronous manner.

The most significant result of *Seaweb '98* is the consistent high quality of received data obtained from

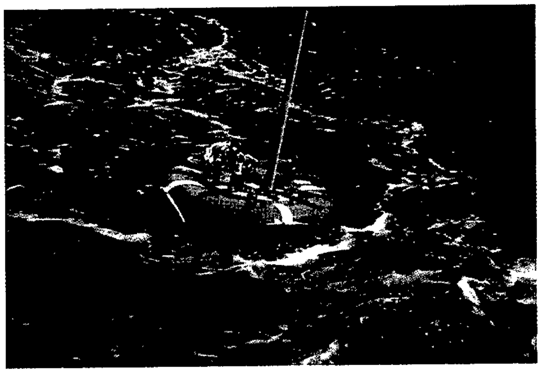


FIGURE 9. In *Seawebs '98*, *'99*, and *2000*, a radio/acoustic communications (racom) buoy provides a very reliable line-of-sight packet-radio link to seaweb servers at the shore command center and on the work boat. The radio link is a 900-MHz spread-spectrum technology commercially known as Freewave. In *Seawebs '99* and *2000*, additional gateway nodes using cellular modems linked via Bell Atlantic and the Internet provide even greater flexibility and provide access by seaweb servers at various locales across the country.

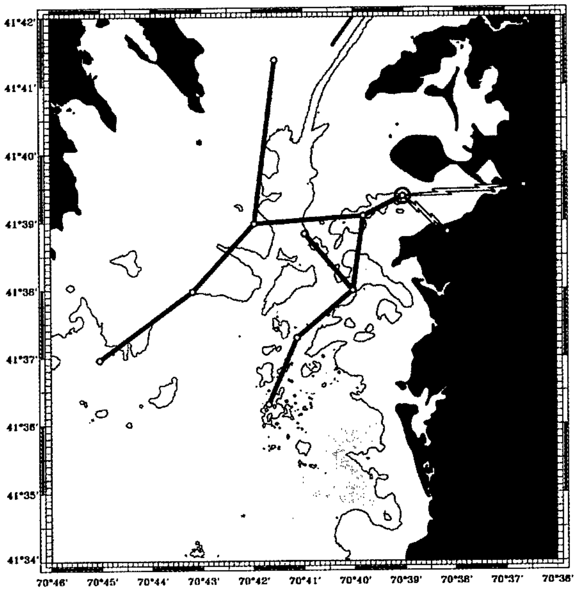


FIGURE 10. *Seaweb '98* demonstrated store and forward of data packets from remote commercial sensors including a CTD, a current vector meter, and a tilt/heading sensor (at the most northerly, westerly, and southerly leaf nodes, respectively) via multiple network links to the racom gateway buoy (large circle). Data packets are then transmitted to the ashore command center via line-of-sight packet radio. An FDMA network with three frequency sets reduced the possibility of packet collisions. Following extensive firmware developments supported by this field testing, the depicted topology was exercised during the final days of the experiment. Isobaths are contoured at 5-m intervals.

remote autonomous sensors. Data packets arrived at the command center via up to four acoustic relays and one RF relay. About 2% of the packets contained major bit-errors attributable to intentional collisions at the master node. The quality of data was very high even after the network was geographically expanded. Reliable direct telesonar communications from the gateway node to a node nearly 7 km distant suggested the network could be expanded considerably more, in spite of the non-ducted 10-m deep channel. The *Seaweb '98* environment could have supported 4-km links using the same ATM875 modems and omnidirectional transducers. Attesting to the channel-tolerant nature of the MFSK modulation, an early phase of testing maintained a 3-km link between two nodes separated by a 1- to 2-m deep rocky shoal. Consistent network degradation occurred during most afternoons and is attributable to summer winds roughening the sea-surface boundary and thus scattering incident acoustic energy. Automated network operations continued during heavy rains and during ship transits through the field.

Seaweb '98 demonstrated the following network concepts: (1) store and forward of data packets; (2) transmit retries and automatic repeat request; (3) packet routing; and (4) cell-like FDMA node grouping to minimize MAI between cells. In addition, the following DADS concepts were demonstrated: (1) networked sensors; (2) wide-area coverage; (3) racom gateway; (4) robustness to shallow-water environment; (5) robustness to shipping noise; (6) low-power node operation with sleep modes; (7) affordability; and (8) remote control. Finally, *Seaweb '98* resulted in dramatic improvements to the ATM875 modem and improved its commercial viability for non-networked applications.

Seaweb '98 observations underscore the differences between acoustic networks and conventional networks. Limited power, low bandwidth, and long propagation times dictate that seaweb networks be simple and efficient. Data compression, forward error correction, and data filtering must be employed at the higher network levels to minimize packet sizes and retransmissions. At the network layer, careful selection of routing is required to minimize transmit energy, latency, and net energy consumption, and to maximize reliability and security. At the physical and MAC layers, adaptive modulation and power control are the keys to maximizing both channel capacity (bit/s) and channel efficiency (bit-km/joule).

SEAWEB '99 EXPERIMENT

Seaweb '99 continued the annual series of telesonar experiments incrementally advancing the state of the art for undersea wireless networks. During a 6-week period, up to 15 telesonar nodes operated in various network configurations in the 5- to 15-m waters of Buzzards Bay. Network topologies provided compound multi-link routes. All links used a rudimentary form of the telesonar handshake protocol featuring an adaptive power-control technique for achieving sufficient but not excessive SNR at the receiver. Handshaking provided the means for resolving packet collisions automatically using retries from the transmitter or automatic-repeat-request (ARQ) packets from the receiver.

The multi-access strategy was a variation of FDMA wherein the six available 20-tone Hadamard words provided six separate FDMA sets, A through F. Rather than clustering the FDMA sets as in *Seaweb '98*, the notion here was to optimally assign FDMA receiver frequencies to the various nodes in an attempt to minimize collisions through spatial separation and the corresponding transmission loss. This approach represents an important step toward network self-configuration and prefigures the future incorporation of secure CDMA spread-spectrum codes to be uniquely assigned to member nodes during the initialization process.

Node-to-node ranging employed a new implementation of a round-trip-travel time measurement algorithm with 0.1-ms resolution linked to the DSP clock rate. Range estimation simply assuming a constant 1500 m/s sound speed was consistently within 5% of GPS-based measurements for all distances and node pairs.

A significant development was the introduction of the seaweb server. It interprets, formats, and routes downlink traffic destined for undersea nodes. On the uplink, it archives information produced by the network, retrieves the information for an operator, and provides database access for client users. The server manages seaweb gateways and member nodes. It monitors, displays, and logs the network status. The server manages the network routing tables and neighbor tables and ensures network interoperability. *Seaweb '99* modem firmware permitted the server to remotely reconfigure routing topologies, a foreshadowing of future self-configuration and dynamic network control. The seaweb server is a graphical set of LabView virtual instruments implemented under Windows NT on a laptop PC. A need for the server was illustrated when operators bypassed server oversight and inadvertently produced a circular routing where a trio of nodes continuously passed a packet between themselves until battery depletion finally silenced the infinite loop.

In *Seaweb '99*, the server simultaneously linked with a Bell Atlantic cellular digital packet data (CDPD) gateway node via the Internet and with the packet-radio racom gateway link via a serial port. A milestone was the establishment of a gateway-to-gateway route through the seaweb server that was exercised automatically over a weekend.

Seaweb '99 included an engineering test for the "Front-Resolving Observation Network with Telemetry" (FRONT) application, with large acoustic Doppler current profiler (ADCP) data packets synthesized and passed through the network with TDMA scheduling. A study of network capacity examined the periodic uplinking of data packets while asynchronously issuing server-generated downlink commands to poll sensors.

For every packet received by a *Seaweb '99* node, the modem appended link metrics such as bit-error rate (BER), automatic gain control (AGC), and SNR. These diagnostics aided post-mortem system analysis. Performance correlated strongly with environmental factors such as refraction, bathymetry, wind, and shipping, although no attempt was made to quantify these relationships in *Seaweb '99*.

The ATM875 second-generation telesonar modem again served as the workhorse modem for all network nodes. During the last phase of the experiment, progress was thwarted by memory limitations of the Texas Instruments TMS320C50 DSP. A firmware bug could not be adequately resolved because of lack of available code space for temporary in-line diagnostics. Consequently, the final days of the test reverted to a prior stable version of the *Seaweb '99* code and the 15-node network charted in Figure 11 covered a less ambitious area than originally intended. These limitations plus the desire to begin implementing FHSS and DSSS signaling motivated the initiation of ATM885 third-generation telesonar modem development for *Seaweb 2000*.

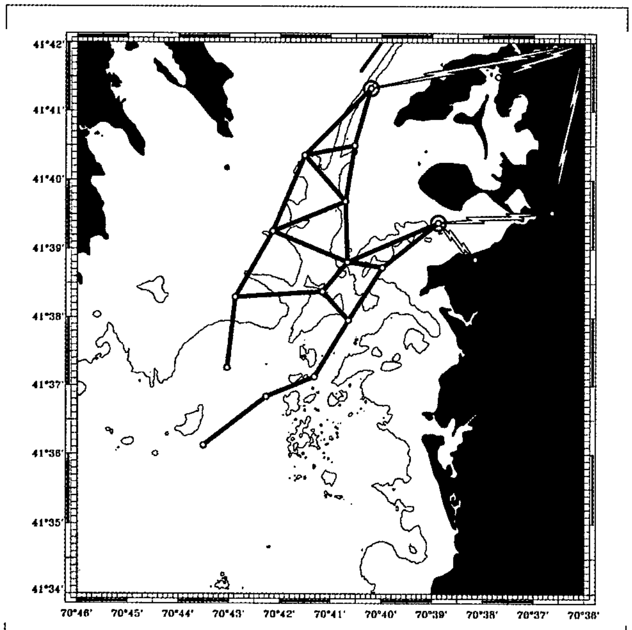


FIGURE 11. *Seaweb '99* explored the use of handshaking and power control. An ADCP sensor node, a tilt/heading sensor node, and a CTD sensor node generated data packets, and the network routed them through various paths. The racom gateway node (easterly large circle) again provided a solid link to shore. A second gateway node (northerly large circle) installed on a Coast Guard caisson near the Bourne canal provided a Bell Atlantic cellular modem link to the Internet and then to the command center. The Seaweb server running on a laptop PC managed both gateway connections and archived all network activity.

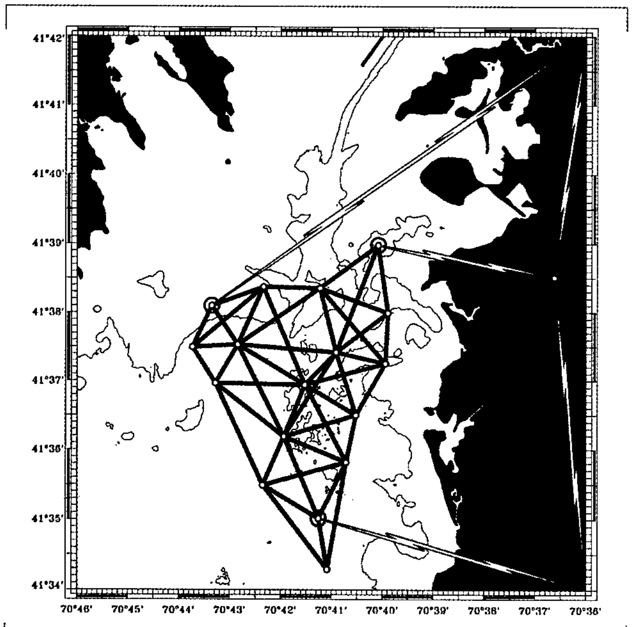


FIGURE 12. *Seaweb 2000* exercised the telesonar handshake protocol in a network context. The 17-node seaweb network delivered oceanographic data from sensor nodes to gateway nodes—one with line-of-sight packet radio and two with cellular telephone modems. During the final week, a seaweb super server operating at the Oceans 2000 Conference in Providence, RI, administered *Seaweb 2000* via the Internet.

SEAWEB 2000 EXPERIMENT

The *Seaweb 2000* network included up to 17 nodes, with one of the fully connected configurations charted in Figure 12. This experiment achieved major advances in both hardware and firmware.

Use of the ATM875 modem during *Seawebs '98* and *'99* continually thwarted progress in firmware development because of limited memory and processing speed. The ATM885 modem shown in Figure 13 overcomes these shortcomings with the incorporation of a more powerful DSP and additional memory. Now, telesonar firmware formerly encoded by necessity as efficient machine language is reprogrammed on the ATM885 as a more structured set of algorithms. The *ForeFRONT-1* (November 1999), *FRONT-1* (December 1999), *ForeFRONT-2* (April 2000), *Sublink 2000* (May 2000), and *FRONT-2* (June 2000) experiments hastened the successful transition of *Seaweb '99* firmware from the ATM875 to the ATM885. These intervening seaweb applications were stepping stones toward achieving basic ATM885 hardware readiness prior to instituting *Seaweb 2000* upgrades.

Seaweb 2000 implements in firmware the core features of a compact, structured protocol. The protocol efficiently maps network-layer and MAC-layer functionality onto a physical layer based on channel-tolerant, 64-bit utility packets and channel-adaptive, arbitrary-length data packets. Seven utility packet types are implemented for *Seaweb 2000.* These packet types permit data transfers and node-to-node ranging. A richer set of available utility packets is being investigated with OPNET simulations, but the seven core utility packets provide substantial networking capability.

The initial handshake consists of the transmitter sending an RTS packet and the receiver replying with a CTS packet. This roundtrip establishes the communications link and probes the channel to gauge optimal transmit power. Future enhancements to the protocol will support a choice of data modulation methods, with selection based on channel estimates derived from the RTS role as a probe signal. A "busy" packet is issued in response to an RTS when the receiver node decides to defer data reception in favor of other traffic. Following a successful RTS/CTS handshake, the data packet(s) are sent. The *Seaweb 2000* core protocol provides for acknowledgments, either positive or negative, of a data message. The choice of acknowledgment type will depend on the traffic patterns associated with a particular network mission. *Seaweb 2000* explored the factors that will guide this application-specific choice.

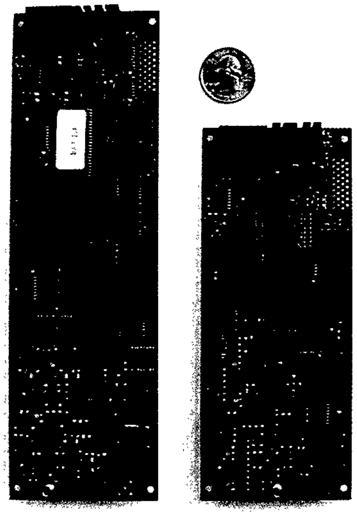


FIGURE 13. The TMS320C5410-based ATM885 telesonar modem debuted in *Seaweb 2000* with a four-fold increase in memory and processing speed over the TMS320C50-based ATM875. This hardware upgrade reduced battery-energy consumption and overcame firmware-development limitations experienced in *Seaweb '99*. The ATM885 supports 100 million instructions per second (MIPS) and 320K words of memory compared with 25 MIPS and 74K available from the ATM875.

A "ping" utility packet initiates node-to-node and node-to-multinode identification and ranging. An "echo" packet is the usual response to a received ping.

In *Seaweb 2000*, FDMA architectures are superseded by hybrid CDMA/TDMA methods for avoiding mutual interference. FDMA methods sacrifice precious bandwidth and prolong the duration of a transmission, often aggravating MAI rather than resisting it. Furthermore, the use of a small number of frequency sets is viewed as an overly restrictive networking solution. Although these drawbacks were expected, *Seawebs '98* and *'99* employed FDMA primarily for ease of implementation as a simple extension to the rigid ATM875 telesonar machine code. The ATM885 permits a break from those restrictions.

Seaweb 2000 execution fully incorporates the experimental approach tried in *Seaweb '99* of establishing two parallel networks—one in air at the command center and one in the waters of Buzzards Bay. This approach minimizes time-consuming field upgrades by providing a convenient network for troubleshooting deployed firmware and testing code changes prior to at-sea downloads.

As a further analysis aid, all modems now include a data-logging feature. All output generated by the ATM885 and normally available via direct serial connection is logged to an internal buffer. Thus, the behavior of autonomous nodes can be studied in great detail after recovery from the sea. To take maximum advantage of this capability, *Seaweb 2000* code includes additional diagnostics related to channel estimation (e.g., SNR, multipath spread, Doppler spread, range rate, etc.), demodulation statistics (e.g., BER, AGC, intermediate decoding results, power level, etc.), and networking (e.g., data packet source, data packet sink, routing path, etc.). For seaweb applications, the data-logging feature can also support the archiving of data until such time that an adjacent node is able to download the data. For example, a designated sink node operating without access to a gateway node can collect all packets forwarded from the network and telemeter them to a command center when interrogated by a gateway (such as a ship arriving on station for just such a data download).

Increasing the value of diagnostic data, the C5410 real-time clock is maintained even during sleep state. Although this clock may not have the stability required for certain future network applications, its availability permits initial development of in-water clock-synchronization techniques.

The new ATM885 modem also includes a provision for a watchdog function hosted aboard a microchip independent of the C5410 DSP. The watchdog resets the C5410 DSP upon detection of supply voltage drops or upon cessation of DSP activity pulses. The watchdog provides a high level of fault tolerance and permits experimental modems to continue functioning in spite of system errors. A watchdog reset triggers the logging of additional diagnostics for thorough troubleshooting after modem recovery.

An aggressive development schedule following *Seaweb '99* and preceding *Seaweb 2000* matured the seaweb server as a graphical user interface with improved reliability and functionality consistent with *Seaweb 2000* upgrades.

Recent telesonar engineering tests have played host to an applied research effort known as SignalEx [30]. This research uses the telesonar test beds to record high-fidelity acoustic receptions and measure relative performance for numerous signaling methods. *Seaweb 2000* hosted SignalEx testing during the second week of testing. The advantage of coupling SignalEx research with seaweb engineering is that both activities benefit—SignalEx gains resources and seaweb gains added empirical test control. By the fifth week, the major *Seaweb 2000* engineering developments reached a level of stability permitting experimental use of acoustic navigation methods for node localization, cost functions for optimized network routing, and statistics gathering for network traffic analysis.

In summary, the specific implementation objectives of *Seaweb 2000* are (1) packet forwarding through network, under control of remotely configurable routing table; (2) 64-bit header; (3) improved software interface between network layer and modem processing; (4) improved wake-up processing, i.e., detection of 2-of-3 or 3-of-4 tones, rather than 3-of-3; (5) improved acquisition signal, i.e., one long chirp, rather than three short chirps; (6) improved channel-estimation diagnostics; (7) logging of channel estimates; (8) RTS/CTS handshaking; (9) configurable enabling of RTS/CTS handshake; (10) power control; (11) watchdog; (12) ARQ; (13) packet time-stamping; and (14) a simple form of adaptive modulation restricted solely to parameter selection for Hadamard MFSK modulation.

The new ATM885 hardware and the *Seaweb 2000* protocols are major strides toward the ultimate goal of a self-configuring, wireless network of autonomous undersea devices.

CONCLUSION

Telesonar is an emerging technology for wireless digital communications in the undersea environment. Telesonar transmission channels include shallow-water environments with node-to-node separations hundreds of times greater than the water depth. Robust, environmentally adaptive acoustic links interconnect undersea assets, integrating them as a unified resource.

Seaweb offers a blueprint for telesonar network infrastructure. Warfare considerations stipulate that the network architecture will support rapid installation, wide-area coverage, long standoff range, invulnerability, and cross-mission interoperability. Seaweb is an information system compatible with low bandwidth, high latency, and variable quality of service. Seaweb connectivity emphasizes reliability, flexibility, affordability, energy efficiency, and transmission security. Network interfaces to manned command centers via gateway nodes such as the racom buoy are an essential aspect of the seaweb concept. C^3N via seaweb supports common situational awareness and collective adaptation to evolving rules of engagement. Seaweb revolutionizes naval warfare by ultimately extending network-centric operations into the undersea battlespace.

The *Seaweb '98*, *'99*, and *2000* experiments incrementally advanced telesonar underwater acoustic signaling and ranging technology for undersea wireless networks. The constraints imposed by acoustic transmission through shallow-water channels have yielded channel-tolerant signaling methods, hybrid multi-user access strategies, novel network topologies, half-duplex handshake protocols, and iterative power-control techniques. *Seawebs '98* and *'99*, respectively, included 10 and 15 battery-powered, anchored telesonar nodes organized as non-centralized, bidirectional networks. These tests demonstrated the feasibility of battery-powered, wide-area undersea networks linked via radio gateway buoy to the terrestrial internet. Testing involved delivery of remotely sensed data from the sea and remote control from manned command centers ashore and afloat. *Seaweb 2000* included 17 nodes equipped with new telesonar modem hardware. It introduced a compact protocol anticipating adaptive network development.

In late summer, *Seaweb 2001* will be conducted in a large expanse of 30- to 300-m waters adjacent to San Diego, CA. The annual seaweb experiments will continue to extend area coverage, resource optimization, network capacity, functionality, and quality of service. Active research includes spread-spectrum signaling, directional transducers [31], *in situ* channel estimation, adaptive modulation, *ad hoc* network initialization, and node ranging and localization.

SSC San Diego is applying seaweb technology for ocean surveillance (DADS Demonstration Project), littoral ASW (Hydra Project), oceanographic research (FRONT Project), submarine communications (Sublink Project) and UUV command and control (SLOCUM and EMATT). Additional applications are proposed.

Deployable autonomous undersea systems will enhance the warfighting effectiveness of submarines, maritime patrol aircraft, amphibious forces, battle groups, and space satellites. Wide-area sensor grids, leave-behind multistatic sonar sources, mine-hunting robots, swimmer-delivery systems, and autonomous vehicles are just a few of the battery-powered, offboard devices that will augment high-value space and naval platforms. Distributed system architectures offer maximum flexibility for addressing a wide array of ocean environments and military missions.

ACKNOWLEDGMENTS

SSC San Diego established the Seaweb Initiative to advance science and technology capabilities for naval command, control, communications, computers, intelligence, surveillance, and reconnaissance (C^4ISR).

Personnel contributing to the Seaweb Initiative are Keyko McDonald, Dave Rees, Dick Shockley, Mark Hatch, Tedd Wright, and Joan Kaina.

The primary sponsor of the Seaweb experiments is the ONR Sensors, Sources, and Arrays Program (ONR 321SS). ONR 321SS also funded the development of telesonar test beds, directional telesonar transducers, and the seaweb server. The SSC San Diego In-house Laboratory Independent Research (ILIR) Program sponsored the telesonar channel measurements and modeling reported here.

Benthos, Inc. personnel Ken Scussel, Dave Porta, John Baker, Jim Hardiman, Dale Green, Jack Crosby, Steve Niland, Bob Burns, Tom Tuite, and Steve Fantone contributed to this work. Benthos' telesonar products and their involvement in seaweb development resulted from the August 1999 acquisition of Datasonics, Inc.

Delphi Communication Systems (DCS) personnel Michael Wolf, Steve Merriam, Ethem Sozer, John Proakis, and Rami Mehio contributed to this work.

DCS and Datasonics involvement in *Seaweb '98* was sponsored by an SBIR topic N93-170 Phase-2 contract administered by Phil DePauk and Linda Whittington of SPAWAR. DCS and Benthos' involvement in *Seawebs '99* and *2000* was sponsored by an SBIR topic N97-106 Phase-2 contract administered by Al Benson, Tom Curtin, Don Davison, Dave Johnson, and Doug Harry of ONR. In addition, Datasonics and Benthos invested independent research and development (IR&D) resources in support of *Seawebs '98* and *'99*.

Dan Codiga, Philip Bogden, Dennis Arbige, and Adam Houk of the University of Connecticut's Department of Marine Sciences contributed to *Seawebs '99* and *2000* and implemented the cellular-modem gateway and the ADCP sensor node. John Newton (Polar Associates, Inc.) performed environmental analysis in support of *Seaweb '98*. Kent Raysin (SAIC) participated in *Seaweb '99*, and Mike Porter (SAIC) participated in the SignalEx portion of *Seaweb 2000*.

AUTHORS

Robert K. Creber

BS in Physics, Florida Institute of Technology, 1984

Current Research: Undersea acoustic modems and networks.

Christopher L. Fletcher

BS in Electrical Engineering, University of Illinois at Champaign/Urbana, 1997

Current Research: Telesonar modems; undersea acoustic networks; lightweight underwater acoustic arrays.

Paul A. Baxley

MS in Oceanography, University of California at San Diego, 1998

Current Research: Underwater acoustic communication modeling; matched-field source localization and tracking; seafloor geoacoustic property inversion.

Kenneth E. Rogers

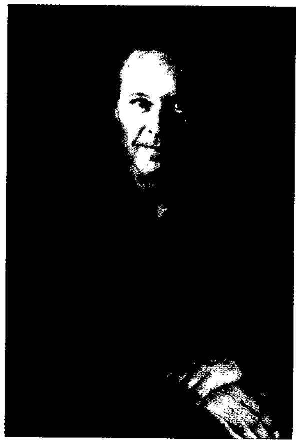
MS in Electrical Engineering, Brigham Young University, 1969

Current Research: Undersea sensors and gateway communications for deployable surveillance systems.

Donald C. Davison

Ph.D. in High Energy Physics, University of California at Riverside, 1969

Current Research: Off-board and deployable surveillance systems.



Joseph A. Rice

M.S. in Electrical Engineering, University of California at San Diego, 1990

Current Research: Ocean sound propagation; sonar systems analysis; undersea wireless networks.

REFERENCES

1. Kilfoyle, D. B. and A. B. Baggeroer. 2000. "The State of the Art in Underwater Acoustic Telemetry," *IEEE Journal of Oceanic Eng*ineering, vol. 25, no. 1, pp. 4–27.
2. Catipovic, J. A., M. Deffenbaugh, L. Freitag, and D. Frye. 1989. "An Acoustic Telemetry System for Deep Ocean Mooring Data Acquisition and Control," *Proceedings of IEEE Oceans '89 Conference*, September.
3. Merriam, S. and D. Porta. 1993. "DSP-Based Acoustic Telemetry Modems," *Sea Technology*, May.
4. Porta, D. 1996. "DSP-Based Acoustic Data Telemetry," *Sea Technology*, February.
5. Rice, J. A. and K. E. Rogers. 1996. "Directions in Littoral Undersea Wireless Telemetry," *Proceedings of The Technical Cooperation Program (TTCP) Symposium on Shallow-Water Undersea Warfare*, vol. 1, pp. 161–172.
6. Scussel, K. F., J. A. Rice, and S. Merriam. 1997. "New MFSK Acoustic Modem for Operation in Adverse Underwater Acoustic Channels," *Proceedings of IEEE Oceans '97 Conference*, Halifax, Nova Scotia, Canada, October, pp. 247-254.
7. Green, M. D. 2000. "New Innovations in Underwater Acoustic Communications," *Proceedings: Oceanology International*, March, Brighton, UK.
8. Jahn, E., M. Hatch, and J. Kaina. 1999. "Fusion of Multi-Sensor Information from an Autonomous Undersea Distributed Field of Sensors," *Proceedings of Fusion '99 Conference*, July, Sunnyvale, CA.
9. McGirr, S., K. Raysin, C. Ivancic, and C. Alspaugh. 1999. "Simulation of Underwater Sensor Networks," *Proceedings of IEEE Oceans '99 Conference*, September, Seattle, WA.
10. Curtin, T. B., J. G. Bellingham, J. Catipovic, and D. Webb. 1993. "Autonomous Oceanographic Sampling Networks," *Oceanography*, vol. 6, pp. 86–94.
11. Sozer, E. M., M. Stojanovic, and J. G. Proakis. 2000. "Underwater Acoustic Networks," *IEEE Journal of Oceanic Eng*ineering, vol. 25, no. 1, pp. 72–83.
12. Rice, J. A. 2000. "Telesonar Signaling and Seaweb Underwater Wireless Networks," *Proceedings of NATO Symposium on New Information Processing Techniques for Military Systems*, 9 to 11 October, Istanbul, Turkey.
13. Rice, J. A. 1997. "Acoustic Signal Dispersion and Distortion by Shallow Undersea Transmission Channels," *Proceedings of NATO SACLANT Undersea Research Centre Conference on High-Frequency Acoustics in Shallow Water*, July, pp. 435–442.
14. Rice, J. A. and R. C. Shockley. 1998. "Battery-Energy Estimates for Telesonar Modems in a Notional Undersea Network," *Proceedings of the MTS Ocean Community Conference*, Marine Technical Society, vol. 2, pp. 1007–1015.
15. Karn, P. 1990. "MACA-A New Channel Access Method for Packet Radio," *Proceedings of the Amateur Radio Relay League/Canadian Radio Relay League (ARRL/CRRL) 9th Computer Network Conference*, September.
16. Rice, J. A. and M. D. Green. 1998. "Adaptive Modulation for Undersea Acoustic Modems," *Proceedings of the MTS Ocean Community Conference*, vol. 2, pp. 850–855.
17. Green, M. D. and J. A. Rice. 2000. "Channel-Tolerant FH-MFSK Acoustic Signaling for Undersea Communications and Networks," *IEEE Journal of Oceanic Engineering*, vol. 25, no. 1, pp. 28–39.
18. Sozer, E. M., J. G. Proakis, M. Stojanovic, J. A. Rice, R. A. Benson, and M. Hatch. 1999. "Direct-Sequence Spread-Spectrum-Based Modem for Underwater Acoustic Communication and Channel Measurements," *Proceedings of IEEE Oceans '99 Conference*, September, Seattle, WA.

19. Stojanovic, M., J. G. Proakis, J. A. Rice, and M. D. Green. 1998. "Spread-Spectrum Methods for Underwater Acoustic Communications," *Proceedings of IEEE Oceans '98 Conference,* September, vol. 2, pp. 650–654.
20. Fruehauf, N. and J. A. Rice. 2000. "System Design Aspects of a Steerable Directional Acoustic Communications Transducer for Autonomous Undersea Systems," *Proceedings of Oceans 2000 Conference*, September, Providence, RI.
21. Rice, J. A., V. K. McDonald, M. D. Green, and D. Porta. 1999. "Adaptive Modulation for Undersea Acoustic Telemetry," *Sea Technology*, vol. 40, no. 5, pp. 29–36.
22. Raysin, K., J. A. Rice, E. Dorman, and S. Matheny. 1999. "Telesonar Network Modeling and Simulation," *Proceedings of IEEE Oceans '99 Conference,* September.
23. Proakis, J. G., M. Stojanovic, and J. A. Rice. 1998. "Design of a Communication Network for Shallow-Water Acoustic Modems," *Proceedings of MTS Ocean Community Conference*, November, vol. 2, pp. 1150–1159.
24. McDonald, V. K., J. A. Rice, M. B. Porter, and P. A. Baxley. 1999. "Performance Measurements of a Diverse Collection of Undersea Acoustic Communication Signals," *Proceedings of IEEE Oceans '99 Conference*, September, Seattle, WA.
25. Codiga, D. L., J. A. Rice, and P. S. Bogden. 2000. "Real-Time Delivery of Subsurface Coastal Circulation Measurements from Distributed Instruments Using Networked Acoustic Modems," *Proceedings of IEEE Oceans 2000 Conference*, September, Providence, RI.
26. Baxley, P. A., H. P. Bucker, and J. A. Rice. 1998. "Shallow-Water Acoustic Communications Channel Modeling Using Three-Dimensional Gaussian Beams," *Proceedings of MTS Ocean Community Conference,* vol. 2, pp. 1022–1026.
27. Green, M. D., J. A. Rice, and S. Merriam. 1998. "Underwater Acoustic Modem Configured for Use in a Local Area Network," *Proceedings of IEEE Oceans '98 Conference*, September, vol. 2, pp. 634–638.
28. Green, M. D., J. A. Rice, and S. Merriam. 1998. "Implementing an Undersea Wireless Network Using COTS Acoustic Modems," *Proceedings of MTS Ocean Community Conference*, vol. 2, pp. 1027–1031.
29. McDonald, V. K. and J. A. Rice. 1999. "Telesonar Testbed Advances in Undersea Wireless Communications," *Sea Technology,* vol. 40, no. 2, pp. 17–23.
30. Porter, M. B., V. K. McDonald, J. A. Rice, and P. A. Baxley. 2000. "Relating the Channel to Acoustic Modem Performance," *Proceedings of the European Conference on Underwater Acoustics*, July, Lyons, France.
31. Butler, A. L., J. L. Butler, W. L. Dalton, and J. A. Rice. 2000. "Multimode Directional Telesonar Transducer," *Proceedings of IEEE Oceans 2000 Conference*, September, Providence, RI.

❖

Shallow-Water Acoustic Communications Channel Modeling Using Three-Dimensional Gaussian Beams

Paul A. Baxley, Homer Bucker, Vincent K. McDonald, and Joseph A. Rice
SSC San Diego

Michael B. Porter
SAIC/Scripps Institution of Oceanography

ABSTRACT

Recent progress in the development of a physics-based numerical propagation model for the virtual transmission of acoustic communication signals in shallow water is presented. The ultimate objective is to provide for the prediction of the output of the quadrature detector (QD, an analog of the discrete Fourier transform) in a time-variant, doubly dispersive, shallow-water channel. Current model development concentrates on the modeling of the QD response in the presence of rough boundaries, reserving inclusion of effects caused by a time-varying sea surface or source/receiver motion to future implementations. Three-dimensional Gaussian beam tracing is used so that out-of-plane reflections from rough surfaces or sloping bathymetry can be adequately modeled. Model predictions of the impulse response for a real shallow-water environment are observed to agree well with measured impulse responses.

INTRODUCTION

Recent innovations in shallow-water undersea surveillance and exploration have necessitated the use of the underwater acoustic medium as the primary means of information exchange. Wireless communication between underwater stations separated in range by as much as 5 km with water depths as low as 10 m may be required. This task is complicated by the inherent spatiotemporal variability of this medium, and the complex nature of multipath arrival of energy for shallow-water environments [1]. Figure 1 illustrates some of the major processes that may affect underwater communication signals.

Multipath spread is caused by refraction governed by the sound-speed profile, reflections from boundaries, and scattering from inhomogeneities. Doppler spread arises from source/receiver motion or the motion of the reflectors and scatterers. These phenomena can significantly disperse and distort the signal as it propagates through the channel. A numerical propagation model that simulates these effects is desired for the systematic study of these phenomena. Such a model would also be a useful tool for environment-dependence assessment, performance prediction, and mission planning of communication systems.

This paper describes an approach for a physics-based model designed to simulate multipath spread and Doppler spread of high-frequency underwater acoustic communication signals. Multipath spread is handled via propagation through a refractive medium, as dictated by the sound-speed profile, and by the modeling of reflection and scattering from arbitrarily rough boundaries. Doppler spread is incorporated via the inclusion of source/receiver motion and sea-surface motion. While other phenomena (water mass fluctuations, scattering from water volume inhomogeneities or bubbles) can be responsible for signal distortions, it is believed that those included are the primary sources of spreading for

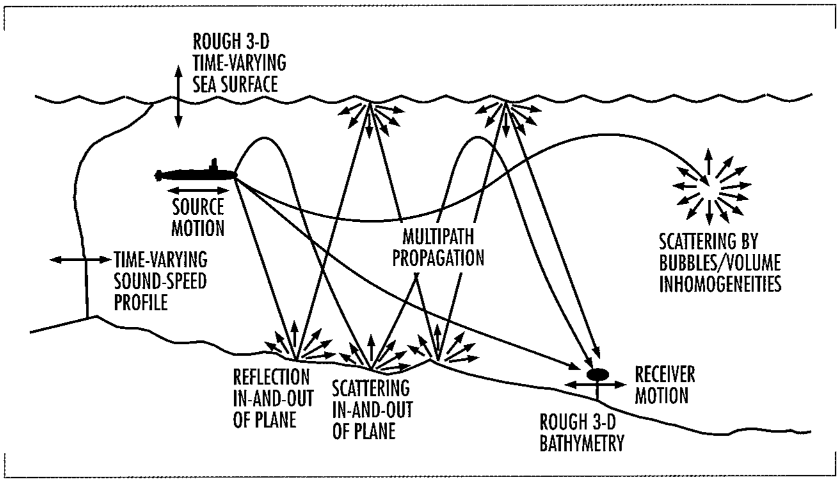


FIGURE 1. Some of the major processes affecting underwater acoustic communications signals.

many realistic problems. The present emphasis is on the modeling of propagation in a bounded refractive medium and the scattering from rough boundaries. The effects of a time-varying sea surface and source/receiver motion are reserved for future implementations.

The basic approach is to model the received output of the quadrature detector (QD) for a transmitted finite-duration constant-wavelength (CW) pulse. The QD is an analog version of the discrete Fourier transform, and provides a convenient means of obtaining the complex Fourier coefficients as a function of time for a finite-duration CW pulse. Because finite-duration CW pulses are common signals in communication schemes, the modeling of such signals is appropriate. However, a broadband QD response (for multiple CW pulses of different frequencies) can also be used to obtain a band-limited impulse response via Fourier synthesis, which is useful for the study of any arbitrary pulse signature.

The pulse is propagated by means of three-dimensional (3-D) Gaussian beams. The consideration of propagation in three dimensions is important because energy can be reflected or scattered in and out of the vertical plane containing both the source and receiver. The high frequencies of communication signals dictate the use of ray-based models over the less-efficient wave models or parabolic-equation approximations. Ray-based models also ensure proper handling of range-dependence and proper reflections from sloping boundaries. The only ray-based method practical for the 3-D problem is Gaussian beams, because the necessity of eigenray determination is eliminated. Ray theory without the use of Gaussian beams requires the determination of eigenrays (rays following paths connecting the source and receiver exactly), which is a formidable task in three dimensions. A dense fan of Gaussian microbeams allows direct modeling of scattering from arbitrarily rough surfaces.

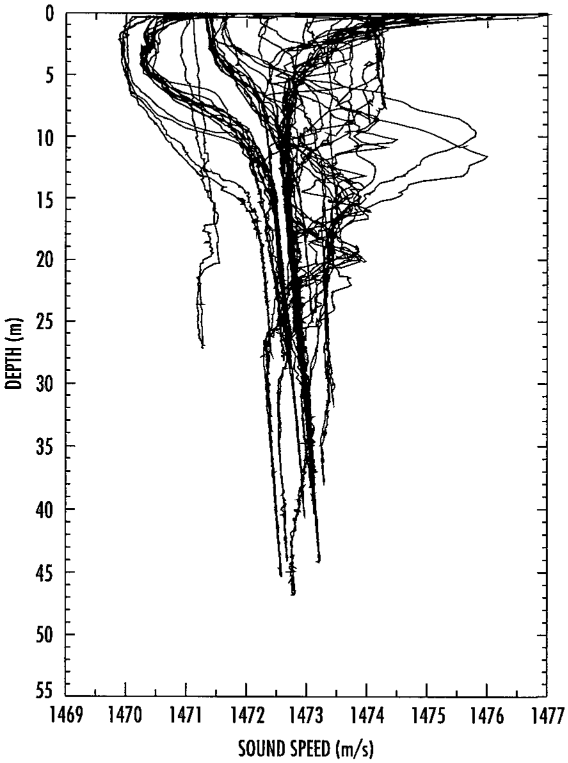


FIGURE 2. Composite of sound-speed profiles measured during the FRONT engineering test.

EXAMPLE OF THE EFFECT OF PROPAGATION CONDITIONS ON COMMUNICATIONS

A compelling example of how ocean channel physics can affect underwater communications was

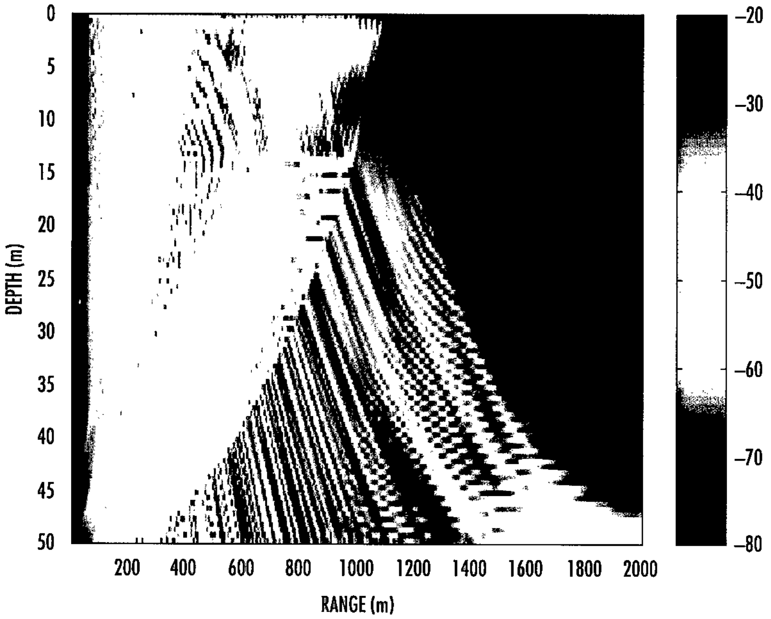


FIGURE 3. Predicted transmission loss during the FRONT engineering test.

provided in engineering tests for the Front-Resolving Observatory with Networked Telemetry (FRONT) oceanographic network. The oceanographic conditions in the area are both interesting and complicated as fresh-river runoff interacts with the tides to generate a persistent front. Figure 2 shows a composite of sound-speed profiles measured at various locations over the duration of the experiment, demonstrating the great variability in the region. Within the upper 10 m of the water column, the channel varies between an upward-refracting (sound speed increases with depth) and downward-refracting (sound speed decreases with depth) channel. Figure 3 shows a transmission loss plot for a typical upward-refracting profile and a communications node (serving as the projector) located on the ocean bottom. This suggests that the influence of sea-surface roughness and time-variability will be greater when the channel is upward refracting.

During the course of the network deployment, there were periods with strong winds followed by relatively calm conditions as the wind speed plot shows in Figure 4A. As the wind speed increases, wave action drives up the ambient noise. At the same time, the roughness of the surface makes it a poor acoustic reflector, so the signal level drops. The combination of the two factors drives the signal-to-noise ratio (SNR) at the bottom-mounted receiver (Figure 4B). This, in turn, drives the overall modem performance as measured by the bit-error rate (Figure 4C). In summary, high winds caused network outages.

This is the simplest of mechanisms driving modem performance. Even with strong SNR, a modem that relies on a tap-delay line for adaptive equalization may fail if the multipath spread becomes too long. Similarly, a modem may fail to track Doppler changes, which is yet another dimension to the parameter space affecting modem performance.

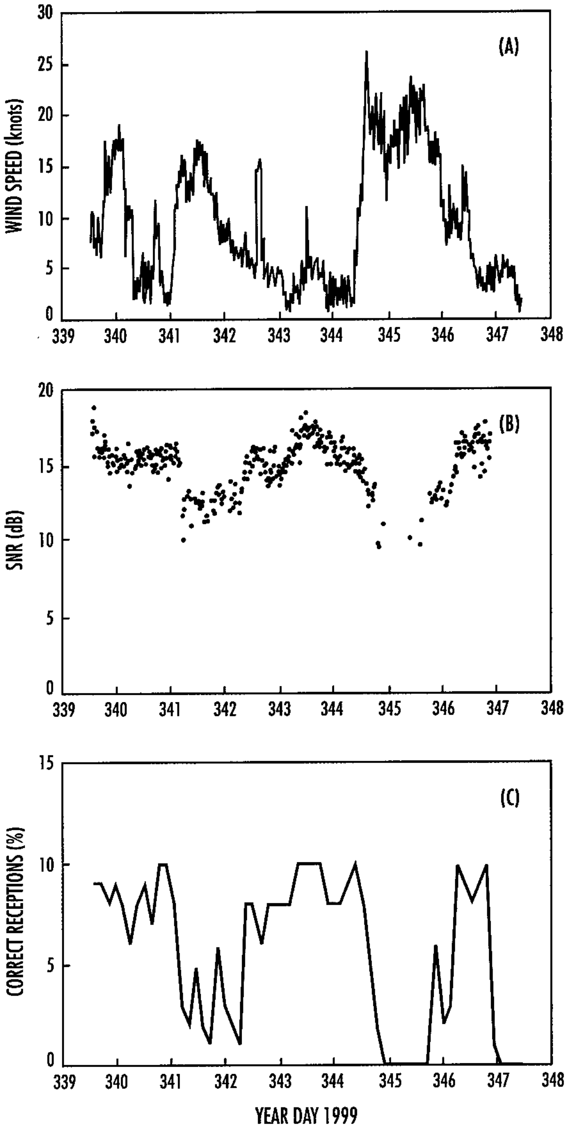


FIGURE 4. (A) During the FRONT network-engineering test, the wind speed varied considerably. (B) The wind speed affects the ambient noise and the surface reflectivity, which both drive the SNR at the receiver. (C) Variations in SNR, in turn, drive the performance of the modem.

3-D GAUSSIAN BEAM PROPAGATION MODEL

The 3-D Gaussian beam model is a modified version of that presented by Bucker [2]. For a specified sound-speed profile and seafloor, beams are traced from a source in three dimensions following the laws of ray acoustics and boundary interactions. Ray theory requires the determination of eigenrays, which can be computationally intensive, particularly in three dimensions. Determination of eigenrays is unnecessary in the Gaussian beam formulation. The sound field at a receiver is obtained by combining the coherent contributions of each beam as determined by the closest point of approach (CPA) of the beam path to the receiver. Consider an arbitrary beam path p that travels from a source S and passes close to a receiver X, as shown schematically in Figure 5A. The point x represents the CPA of the beam to the receiver and ρ is the CPA distance. The actual path length (arc length) from

S to x is designated S_x, and is shown linearized in Figure 5B. The pressure at the receiver X associated with this beam path is then given by

$$p = C_n B[\exp(-a\theta^2 + i\omega t)]/S_x, \qquad (1)$$

where C_n is a normalization constant, a is an empirical constant, $\theta = \tan^{-1}(\rho / S_x)$, ω is the angular frequency, and t is the travel time to point x. A spherical wave-front correction equal to $(L-S_x)/c_x$, where $L^2=S_x^2+\rho^2$ and c_x is the sound speed at x, is included in the travel time t. B accounts for energy loss and phase shifts resulting from surface and bottom reflections. See [2] for a fuller explanation of the constants a and C_n.

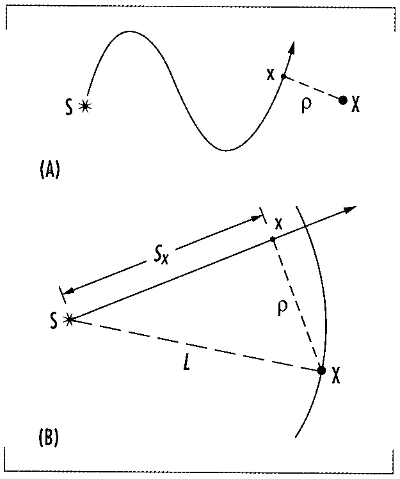


FIGURE 5. Geometry used to determine pressure contribution of a Gaussian beam at CPA to sensor X.

An important feature of the 3-D Gaussian beam model is that the bottom can be specified arbitrarily. Bottom depth data z are specified digitally by the user as a function of the horizontal coordinate directions x and y. Third-order smoothing polynomials are fitted to the bottom data in both directions so that the depth z and the unit normal $\tilde{n}$ can be determined for any arbitrary value of x and y (see appendix B of [2] for details). Bottom interactions are modeled via the specification of the reflection coefficient, or via the calculation of the reflection coefficient from specified geoacoustic parameters (sediment compressional and shear sound speed and attenuation and density). The bottom displacement technique of Zhang and Tindle [3] may also be used. Currently, only a semi-infinite representation for the bottom is implemented; this is sufficient for the high frequencies of interest in communication systems.

Arbitrary specification of the bottom depth implies that scattering problems may be handled directly and deterministically without the use of statistical techniques or approaches only applicable to particular classes of problems because of their underlying assumptions. Because the roughness can be arbitrarily specified, scattering effects are modeled by tracing a dense fan of very fine microbeams, which follow the physics of the interface interactions directly. By this means, problems at shallow grazing angles, such as self-shadowing effects, can be treated. In addition, using an arbitrary specification of bottom depth with 3-D Gaussian beams allows examination of environments possessing significant range-dependence.

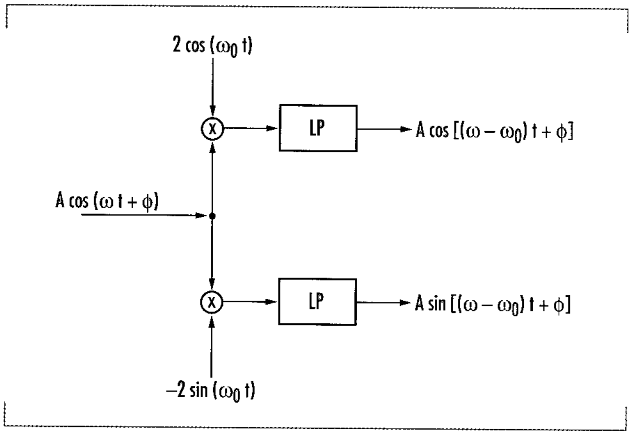


FIGURE 6. Quadrature detector algorithm for a continuous sinusoidal signal.

QUADRATURE DETECTOR RESPONSE

Assume that a continuous sinusoidal signal $A\cos(\omega t+\phi)$ arrives at a receiver, where ω is the frequency, t is the time, and ϕ is a phase shift associated with boundary interactions. If the signal is processed by a quadrature detector, as diagramed in Figure 6, the signal is split with one part being multiplied by $2\cos(\omega_0 t)$ and the other part being multiplied by $-2\sin(\omega_0 t)$. ω_0 is the reference frequency, which should be approximately equal to ω. Both parts are then passed through a low-pass filter to obtain the quadrature components $A\cos[(\omega-\omega_0)t+\phi]$ and $A\sin[(\omega-\omega_0)t+\phi]$, respectively. The complex output R_{QD} of the QD is then simply

$$R_{QD} = A\exp[i(\omega - \omega_0)t + \phi]\,, \qquad (2)$$

This is basically an analog version of the discrete Fourier transform. If $\omega_0 \neq \omega$, R_{QD} will experience a rotation of ω–ω_0 radians per second.

Now assume that the incoming signal is a finite sinusoidal pulse of duration τ seconds and frequency ω. Assume also that the travel time from the source to the receiver along path p is t_p and that the time constant t_c (effective integration time) of the low-pass filters in the QD is τ seconds. For this case, R_{QD} is modulated by a triangle function $T(t)$ that is zero for $t<t_p$, increases linearly from zero to a value of unity at $t=t_p+\tau$, and then decreases linearly back to zero at $t=t_p+2\tau$. Therefore, the quadrature response for a beam travelling along path p is

$$R_{QD} = AT(t)\exp[i(\omega - \omega_0)t + \phi] , \tag{3}$$

If the time constant t_c is larger than the pulse duration τ, R_{QD} remains at the value of the apex until $t>t_p+t_c$. In either case, R_{QD} still experiences the rotation $\omega - \omega_0$ if $\omega_0 \neq \omega$. The only way that ω_0 cannot equal ω in the above scenario is for a Doppler shift to have occurred somewhere along the path p. Therefore, source/receiver motion or sea-surface motion results in a rotation of R_{QD} for a path influenced by that motion. The total quadrature response of a received signal is therefore easily obtained by combining the quadrature responses of all paths contributing to the pressure at the receiver. This is facilitated via the use of the 3-D Gaussian beam model to propagate the energy. Closely spaced microbeams are launched from the source and traced through the refractive, bounded medium. Travel times, phase shifts associated with boundary reflections, and Doppler shifts associated with a moving source/receiver or with a moving sea surface are accumulated for each microbeam as it propagates. This information is then used with Eq. (3) to determine the QD response for each microbeam. A superposition of the QD response for all microbeams then provides the total QD response. If the pulse length is small, the QD response represents an estimate of the channel impulse response. The multipath structure resulting from refraction and the complex interactions of the many microbeams with the rough surface will combine to yield the effect of multipath spread on the QD impulse response. The Doppler shifts accumulated for each microbeam will combine to yield the effect of Doppler spread on the QD impulse response.

MODEL DEMONSTRATION

The environment selected for demonstrating the usefulness of the channel model was that of the SignalEx-99 experiment conducted in April 1999 in a shallow-water (~200 m) region, 6 km southwest of San Diego. Sponsored by the Office of Naval Research, SignalEx-99 was the first in a series of experiments intended to relate channel propagation characteristics to the performance of underwater acoustic communication systems. A detailed description of the experiment is provided by McDonald et al. [4].

The data considered here were linear frequency-modulated (LFM) chirps emitted/received from a source/receiver deployed at a depth of 30 m and source/receiver mounted 6.7 m above the seafloor. The source/receiver systems were telesonar test beds [5 and 6], autonomous units consisting of a single-board computer with a projector and a four-phone vertical line array. The 30-m test bed was deployed from a freely drifting ship, resulting in measurements as a function of time along a fairly constant track. The water depth at the receiver was 210 m, while the water depth decreased in a near linear fashion along the track to a depth of approximately 170 m at a range of 3.8 km from the receiver. Transmissions were made in both directions between the two test beds.

Figure 7 shows the bathymetry and track of the drifting source. Northerly winds caused the ship to drift from a range of about 0 to 4 km (Drift 1). As the range was becoming large and the ship began drifting off the isobath, the ship was repositioned back at a range of about 2 km and allowed to drift again (Drift 2). This conveniently provided a look at the consistency of the Drift 1 results. Once again, the ship drifted to a range of about 6.5 km and was repositioned and moored at a range of 4.75 km providing a look at the stability of the signaling schemes with fixed source-receiver geometry.

Figure 8 shows a typical sound-speed profile measured during the SignalEx-99 experiment. The profile is strongly downward refracting with approximately a 20-m mixed layer at the surface and a slight duct forming near the bottom. It has been determined previously [7] that the bottom in this region may be treated as a fluid with a compressional sound speed of 1572.37 m/s, a compressional attenuation of 0.20 dB/kmHz, and a density of 1.76 g/cm^3.

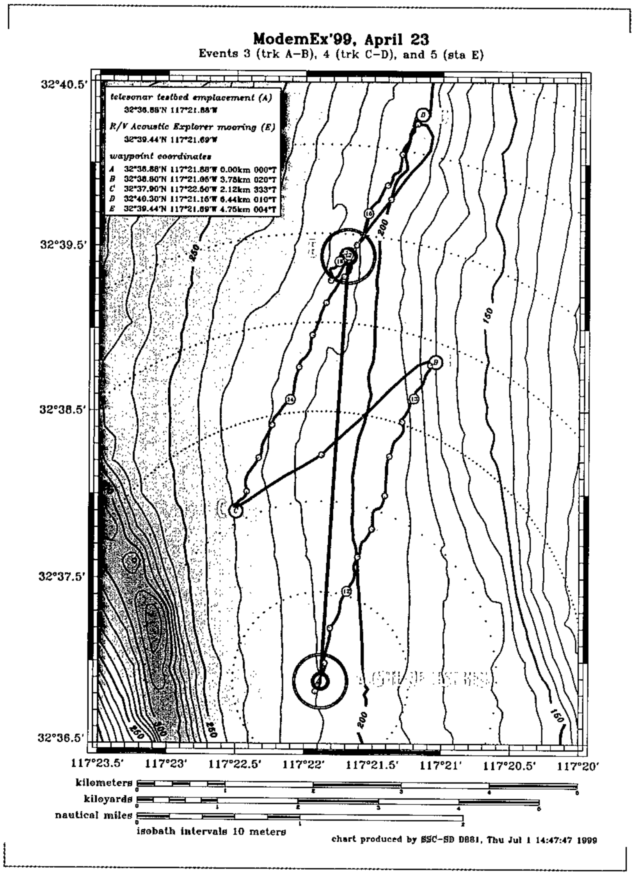

ModemEx'99, April 23
Events 3 (trk A–B), 4 (trk C–D), and 5 (sta E)
telesonar testbed emplacement (A)
32°36.88'N 117°21.88'W
R/V Acoustic Explorer mooring (E)
32°39.44'N 117°21.69'W
waypoint coordinates
A 32°36.88'N 117°21.88'W 0.00km 000°T
B 32°38.80'N 117°21.05'W 3.75km 020°T
C 32°37.90'N 117°22.50'W 2.12km 333°T
D 32°40.30'N 117°21.16'W 6.44km 010°T
E 32°39.44'N 117°21.69'W 4.75km 004°T
32°40.5'
32°39.5'
32°38.5'
32°37.5'
32°36.5'
117°23.5'
117°23'
117°22.5'
117°22'
117°21.5'
117°21'
117°20.5'
117°20'
250
300
250
200
200
160
kilometers
0
1
2
3
4
5
kiloyards
0
1
2
3
4
5
nautical miles
0
1
2
isobath intervals 10 meters
chart produced by SSC-SD D881, Thu Jul 1 14:47:47 1999

FIGURE 7. Bathymetry and source track for SignalEx-99 experiment. Drift 1 is from A to B (source range = 0 to 3.8 km). Drift 2 is from C to D (source range = 2.1 to 6.5 km). The moored station is at E (source range = 4.75 km).

The LFM chirps were transmitted sweeping the 8- to 16-kHz band over a 1-second period. Sixteen chirps were transmitted in 10-minute time frames over a 5-hour period. The direction of the transmission was switched for consecutive 10-minute periods. Theoretically, the impulse response is a combination of these chirps delayed in time according to their path length and attenuated according to volume absorption and reflection loss at the boundaries. The impulse response can be estimated experimentally by correlating the received pulses with a replica of the original transmitted pulse. This produces a sequence of impulses corresponding to each echo in the received waveform, thereby providing a visualization of the impulse response. Figure 9 shows the result of performing this correlation as a function of time. The variation in the multipath structure throughout the experiment is clearly observed. Because absolute times were not available, the first significant peak in each reception was detected and used to provide a leading-edge alignment. Note also that this plot is a composite of the transmissions that alternated between the ship and bottom-mounted test bed.

Figure 10 compares the measured impulse response at a time of day of 12.5 hours (ping number 34) to a simluated response obtained via the 3-D Gaussian beam, quadrature detector model for the same source-receiver configuration. The source range at this time was 2.2 km. The simulation was performed assuming a flat bottom at a water depth of 210 m, and ignoring the effects of rough-surface scattering and time-variability. The measured impulse response has been normalized relative to the maximum, and the arbitrary time scale has been shifted to facilitate comparison with the modeled result. Note that the predicted arrivals agree well with the measured arrivals, indicating that refraction and reflection

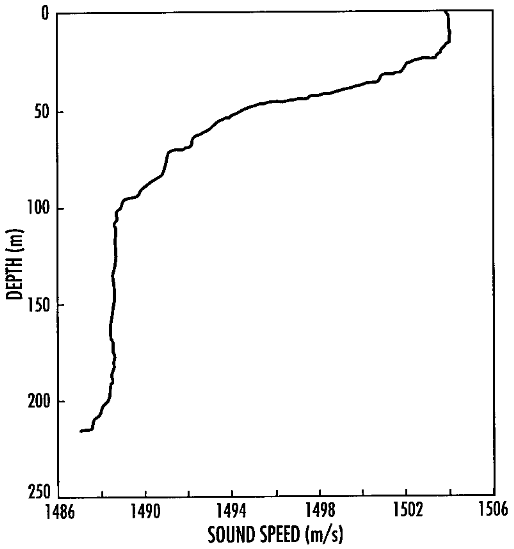


FIGURE 8. Typical sound-speed profile measured during SignalEx-99 experiment.

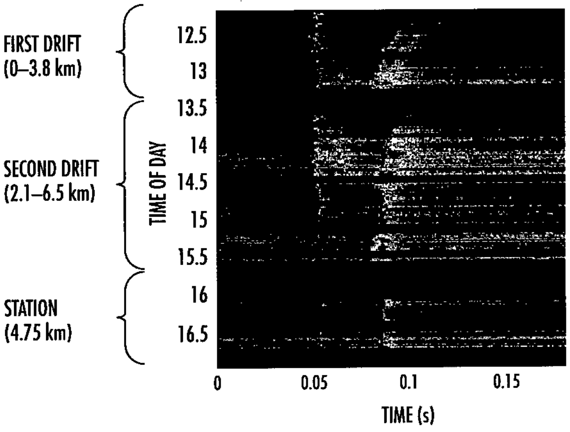


FIGURE 9. Replica correlogram from chirps during SignalEx-99 experiment.

from boundaries are well modeled. Time dicrepancies between arrival paths may be caused by the neglect of the varying bathymetry or errors in the assumed sound-speed profile. The higher resolution of the model results indicates that the first arrival is actually a combination of several arrivals: namely, the direct path, the one-bottom-reflected path, the one-surface-reflected path, and the one-surface-reflected/one-bottom-reflected path. Likewise, the later arrivals are actually a combination of several higher order paths. Note also that the data exhibits a gradual rolloff after the arrival of the pulses, suggesting a reverberant environment. The likely cause of this behavior is the scattering of energy in three dimensions caused by the interaction of rays with the boundaries. Future work will attempt to model these interactions.

Future developments of the model will focus on determining how scattering from rough surfaces, source/receiver motion, and sea-surface motion will influence these responses.

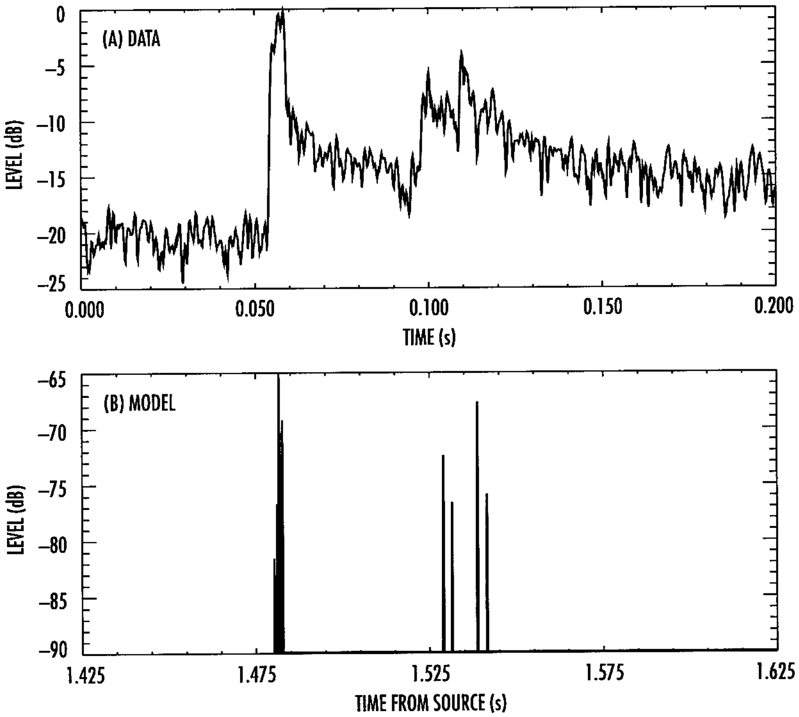


FIGURE 10. Comparison between (A) measured and (B) modeled impulse responses. Time of day = 12.5 hours. Source range = 2.2 km.

MODELING SURFACE AND BOTTOM SCATTER

Scattering from rough boundaries produces losses in signal energy. These losses are two-fold. First, scatter converts energy to higher angles eventually allowing it to penetrate the bottom where it is absorbed. Second, it destroys the coherence of the wave producing what might be termed an apparent loss. For instance, a moving surface will stretch and compress a sinewave reflected from it. If the reflected energy is detected by a matched-filter expecting a perfect sinewave, it will see a reduced power level. This discussion applies, for instance, to a single tone in an M-ary Frequency-Shift Keying (MFSK) signaling scheme, where the tone is detected by a filter bank. If we have a rough bottom with a static geometry, this loss of coherence does not occur. However, if the source or receiver moves, we have a dynamic situation similar to the surface loss just described.

In round numbers, a typical communications carrier gives a wavelength around 10 cm. A classical measure of the role of roughness—the Rayleigh roughness parameter—is the ratio of the roughness to the wavelength (or more precisely, the vertical component of the wavelength). As this number becomes close to unity, losses per bounce become large, perhaps 10 dB, and many of the standard scatter models that assume small roughness fail. The point of this discussion is that 10-cm roughness is easily attained on both surface and bottom boundaries in real environments, implying large boundary losses. Furthermore, the roughness is typically not known to within 10 cm, implying large uncertainty in those same losses and in the resulting transmission loss.

Finally, the actual scatter mechanisms are complicated. In some cases, the air–water interface is the scatterer. In other cases, the bubbles below are likely to be dominant. Similarly, at the ocean bottom, scatter can occur at the interface or by inhomogeneities just below the interface (though not too far below because volume attenuation limits the sediment penetration significantly).

As a first step toward modeling scattering effects, we assume that the boundary roughness dominates the problem, and concentrate first on the bottom roughness. A common approach [8] to characterize this roughness is to use the spatial power spectral density, i.e., the power spectrum of the bottom roughness. Various forms may be used; however, one popular choice is $\Phi(k) \propto k^{-b}$, where b is a measured parameter for the particular site. Suggested values for b are given in [8] along with the RMS roughness that defines the overall amplitude of the spectrum.

Given the spatial power spectral density, we can construct individual realizations of the bottom by using a standard technique. In particular, we convert the power spectrum to an amplitude spectrum by taking its square root. We discretely sample the amplitude and then introduce a random phase. Finally, we do a fast Fourier transform (FFT) to produce and add in the mean depth to obtain a single realization of the bottom. In equations:

$$D(r) = \int A(k) e^{\theta} e^{ikr} dk \tag{4}$$

where $A(k) = \sqrt{\Phi(k)}$ and $\Phi(k) = 5.5 \times 10^{-5} k^{-2.25}$
is the spatial power density spectrum for a particular area.

As a specific example, Figure 11 shows a single realization of the bottom depth using the above described power spectrum. Figure 12 compares the predicted transmission loss for the SignalEx-99 environment using this rough bottom (Figure 12A) with that predicted using a smooth bottom (Figure 12B). The transmission loss calculation was done using the BELLHOP ray/beam model [9], which is a two-dimensional (2-D) version of the 3-D Gaussian beam model. The prediction is for the case of a test bed deployed on the seafloor. Note the fill-in of the shadow zone near the surface at a range of 2 km. There are also changes in the Lloyd-mirror pattern emanating from the source.

The root-mean-square (RMS) height used here is 0.23 m, which is fairly low. There will also be surface scatter that may also be expected to have a larger RMS roughness.

SUMMARY AND FUTURE APPLICATIONS

The channel model outlined in this paper is being developed to aid in the analysis of future underwater acoustic communication systems. The modeling of the QD response via the use of 3-D Gaussian beams enables the inclusion of physical phenomena known to influence such systems in a computationally efficient manner. The model will provide a useful tool for examining the effect of multipath and Doppler spread on the performance of these systems. Because it is designed for use with finite-duration CW pulses, the model can be used directly for the analysis of MFSK systems. Otherwise, the model can also be run for multiple frequencies to obtain a band-limited impulse response via Fourier synthesis.

The following future work is planned. An analytical model will be implemented for a sinusoidally

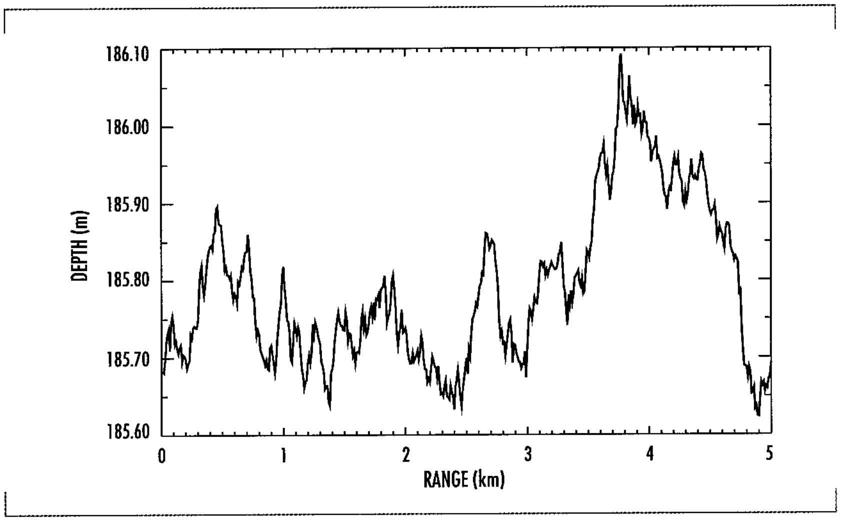


FIGURE 11. Single realization of the bottom depth using a power-law spectrum.

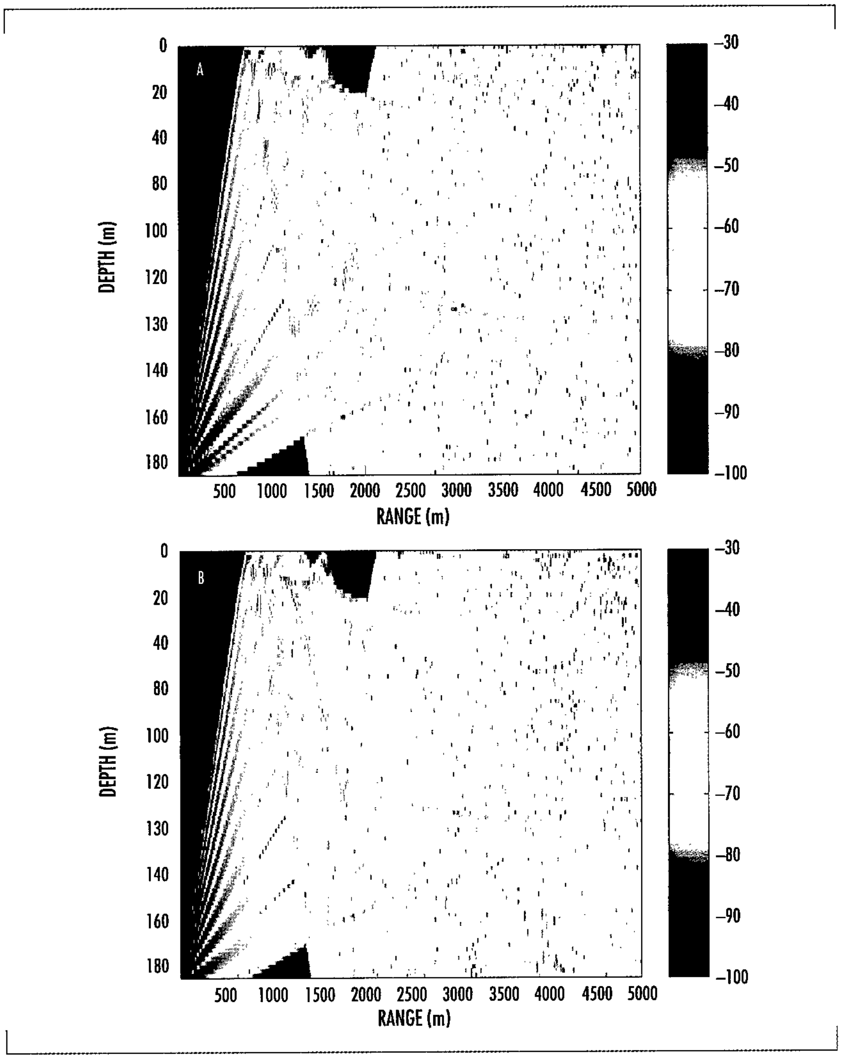


FIGURE 12. Comparison of predicted transmission loss for the SignalEx-99 environment (A) using the rough bottom in Figure 11, and (B) using a smooth bottom. Source is near the seafloor.

corrugated interface and used as a benchmark solution to validate the above ray/beam results for scatter from a rough surface. The 2-D scattering approach will be expanded to provide 3-D scattering for inclusion in the 3-D Gaussian beam quadrature detector model. The importance of 3-D scattering effects on the impulse response will then be studied. These tasks are geared to evaluating the mean energy level in a situation where the processing time is short enough that a "frozen ocean" model is appropriate. Once this is accomplished, a time-varying sea surface will be implemented.

Work is underway to use the model to infer channel characteristics such as coherence time, coherence bandwidth, multipath spread, and Doppler spread. It will also be used as part of a statistically governed Markov process to produce a time-dependent simulation.

ACKNOWLEDGMENTS

The Office of Naval Research (ONR) via the SSC San Diego In-house Laboratory Independent Research (ILIR) Program sponsored this work. The SignalEx-99 experiment was sponsored by ONR 322OM via the Telesonar Signaling Measurements Project. The FRONT program is sponsored by the National Oceanographic Partnership Program (NOPP). The authors also thank Robert Creber and Christopher Fletcher for their help in obtaining experimental data.

AUTHORS

Homer Bucker

Ph.D. in Physics, University of Oklahoma, 1962

Current Research: Matched-field tracking; underwater acoustic propagation modeling using three-dimensional Gaussian beams.

Vincent K. McDonald

BS in Mathematics, San Diego State University, 1988

Current Research: Underwater acoustic communications research; underwater surveillance system design.

Joseph A. Rice

MS in Electrical Engineering, University of California at San Diego, 1990

Current Research: Ocean sound propagation; sonar systems analysis; undersea wireless networks.

Michael B. Porter

Ph.D. in Engineering Sciences and Applied Mathematics, Northwestern University, 1984

Current Research: Array signal processing; acoustic communications; wave propagation.

REFERENCES

1. Rice, J. A. 1997. "Acoustic Signal Dispersion and Distortion by Shallow Undersea Transmission Channels," *Proceedings of the NATO SACLANT Undersea Research Centre Conference on High-Frequency Acoustics in Shallow Water*, pp. 425–442.
2. Bucker, H. P. 1994. "A Simple 3-D Gaussian Beam Sound Propagation Model for Shallow Water," *Journal of Acoustical Society of America*, vol. 95, no. 5, pp. 2437–2440.

3. Zhang, Z. Y. and C. T. Tindle. 1993. "Complex Effective Depth of the Ocean Bottom," *Journal of Acoustical Society of America*, vol. 93, no. 1, pp. 205–213.

4. McDonald, V. K., J. A. Rice, M. B. Porter and P. A. Baxley. 1999. "Performance Measurements of a Diverse Collection of Undersea Acoustic Communication Signals," *Proceedings of IEEE Oceans '99 Conference*, 13 to 16 September, Seattle, WA.

5. McDonald, V. K., and J. A. Rice. 1999. "Telesonar Testbed—Advances in Undersea Wireless Communications," *Sea Technology*, vol. 40, no. 2, pp. 17–23.

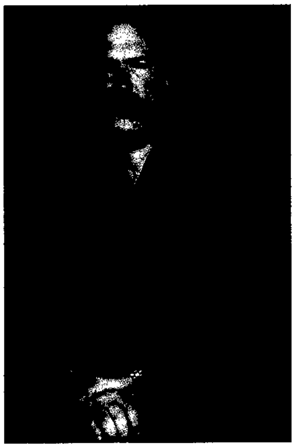
6. McDonald, V. K., J. A. Rice, and C. L. Fletcher. 1998. "An Underwater Communication Testbed for Telesonar RDT&E," *Proceedings of MTS Ocean Community Conference '98*, 16 to 19 November, Baltimore, MD.

7. Baxley, P. A., N. O. Booth, and W. S. Hodgkiss. 2000. "Matched-Field Replica Model Optimization and Bottom Property Inversion in Shallow Water," *Journal of Acoustical Society of America*, vol. 107, no. 3, pp. 1301–1323.

8. Medwin, H. and C. S. Clay. 1998. *Fundamentals of Acoustical Oceanography*, Academic Press, San Diego, CA.

9. Porter, M. B. and Y-C Liu. 1994. "Finite-Element Ray Tracing," *Theoretical and Computational Acoustics*, (D. Lee and M. H. Schultz, eds.) World Scientific Publishing Company, River Edge, NJ, vol. 2, pp. 947–956.

❖



Paul Baxley
MS in Oceanography, Scripps Institution of Oceanography, University of California at San Diego, 1998
Current Research: Underwater acoustic communication modeling; matched-field source localization and tracking; seafloor geoacoustic property inversion.

Advanced Refractive Effects Prediction System (AREPS)

Wayne L. Patterson
SSC San Diego

ABSTRACT

In 1987, SSC San Diego fielded the Integrated Refractive Effects Prediction System (IREPS), the world's first electromagnetic prediction system for shipboard use. Advances in research and technology have led to the replacement of IREPS with the Advanced Refractive Effects Prediction System (AREPS). AREPS computes and displays radar probability of detection, propagation loss and signal-to-noise ratios, electronic-support-measures vulnerability, UHF/VHF communications, and surface-borne surface-search radar capability vs. range, height, and bearing from the transmitter.

INTRODUCTION

In 1987, SSC San Diego provided the U.S. Navy's operational fleet with its first capability to assess the effects of the atmosphere on the performance of electromagnetic (EM) systems such as radars and radios. This assessment system was named the Integrated Refractive Effects Prediction System (IREPS). IREPS was hosted on the Hewlett-Packard 9845 desktop calculator. The EM propagation models of IREPS were semi-empirical and assumed that the atmosphere is homogeneous in the horizontal. IREPS also assumed the earth's surface was water. As desktop computing developed and EM propagation modeling advanced, the various assumptions of IREPS were overcome. In response to a request from Commander, Sixth Fleet during the Bosnian campaign, a new assessment system, the Advanced Refractive Effects Prediction System (AREPS) was fielded for fleet operations.

AREPS computes and displays radar probability of detection, propagation loss and signal-to-noise ratios, electronic-support-measures (ESM) vulnerability, UHF/VHF communications, and surface-borne surface-search radar capability vs. range, height, and bearing from the transmitter.

The power of AREPS derives from its Windows 95/NT interface, making full use of pop-up menus, object linking and embedding (OLE) features such as file drag and drop and graphics export, and extensive online help with color graphic examples.

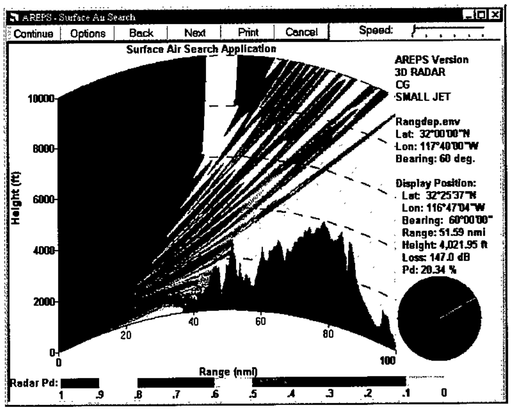
At the core of AREPS is our Advanced Propagation Model (APM), a hybrid ray-optic and parabolic equation (PE) model that uses the complementary strengths of both methods to construct a fast yet very accurate composite model. Depending on the requirements of the tactical decision aid, APM will run in several different modes. For the full hybrid mode, APM is much faster than PE models alone, with overall accuracy at least as good as the pure PE models. With its airborne submodel, APM can solve problems for very high elevation angles where PE methods would not normally be used.

APM allows for range-dependent refractivity over various sea and/or terrain paths. Not only does the terrain path include variable terrain heights, it may also include range-varying dielectric ground constants for finite conductivity and vertical polarization calculations. APM considers absorption of electromagnetic energy by oxygen and water vapor. APM accounts for all normal propagation mechanisms, including troposcatter and the anomalous propagation mechanisms of subrefraction, superrefraction, and ducting.

AREPS DISPLAYS

The primary AREPS displays are height vs. range and bearing coverage and path loss vs. height/range and bearing. Figure 1 shows such a coverage display for shipborne air-search radar with its probability of detecting a "small-sized" jet. For this case, the atmosphere is range-dependent, with a surface-based duct existing at the transmitter location, rising to become an elevated duct over the terrain features. To the lower right of the coverage display is a small map, in a simulated plan-position-indicator (ppi) picture format, showing the transmitter location, the display's current bearing, and the terrain heights.

At the top of the display window is a series of buttons that allow you to animate the display in bearing, both forward and backward, to pause the animation, and to obtain a printed copy of the display. Because AREPS is a Windows 95/NT program, the full capabilities of the operating system are available. For example, should you desire to brief the display, you may "copy" the display to the Windows 95/NT clipboard and "paste" it directly into a presentation package such as Microsoft PowerPoint. To obtain loss vs. range and or height displays (Figures 2 and 3), you simply click the right mouse button on the coverage display.



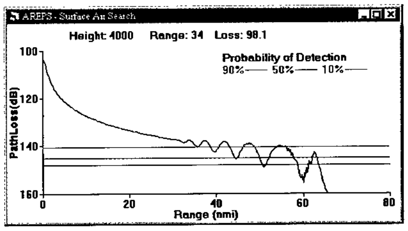
AREPS Version
3D RADAR
CG
SMALL JET

Rangdep.env
Lat: 32°00'00"N
Lon: 117°40'00"W
Bearing: 60 deg.

Display Position:
Lat: 32°25'37"N
Lon: 116°47'04"W
Bearing: 60°00'00"
Range: 51.59 nmi
Height: 4,021.95 ft
Loss: 147.0 dB
Pd: 20.34 %

FIGURE 1. AREPS radar probability of detection coverage display.

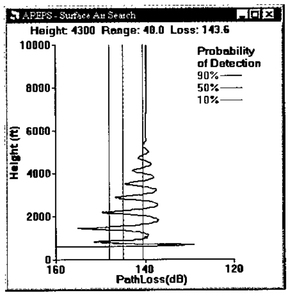
Figure 4 shows the coverage for an airborne transmitter in the presence of an elevated duct; Figure 5 shows the simultaneous surface-based radar coverage and ESM vulnerability; and Figure 6 shows the UHF communication assessment. Note also the three earth surface depictions: dual curved, curved, and flat.



Height: 4000 Range: 34 Loss: 98.1

FIGURE 2. AREPS loss vs. range display.

In addition to coverage displays, the effects of radar cross section variability as a function of viewing angle, ship displacement, ship height, and range are combined with the APM capabilities of range-dependent environments and terrain to produce a bar graph display (Figure 7) of detection for five classes of ship targets. These classes range from small (a patrol boat) to a very large warship (aircraft carrier). The viewing angle variability is displayed as subbars within each ship class. These angles are labeled minimum, maximum, and average, corresponding to bow, beam, and quarter.

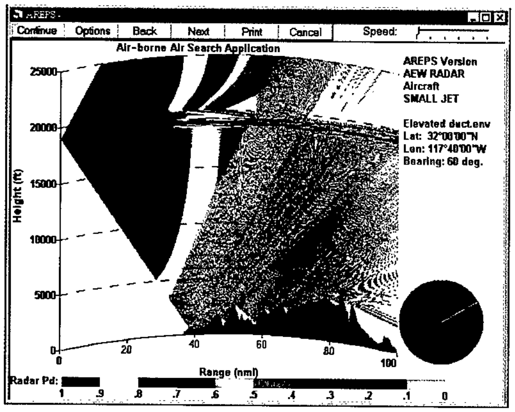


Height: 4300 Range: 40.0 Loss: 143.6

FIGURE 3. AREPS loss vs. height display.

EM Systems Database

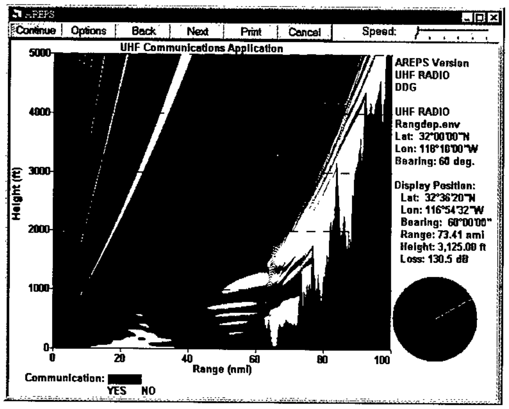
AREPS is an unclassified program and, as such, does not include a preestablished EM system parameter database. Users are solely responsible for creating a system parameter database appropriate to their situation. To assist in this task, a database creation and maintenance capability is provided that uses fill-in-the-blank forms. Figure 8 shows such a form



AREPS Version
AEW RADAR
Aircraft
SMALL JET

Elevated duct.env
Lat: 32°00'00"N
Lon: 117°40'00"W
Bearing: 60 deg.

FIGURE 4. AREPS airborne air-search application.

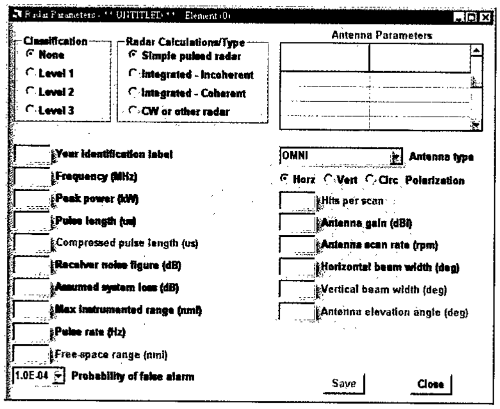


AREPS Version
UHF RADIO
DDG

UHF RADIO
Rangdep.env
Lat: 32°00'00"N
Lon: 116°16'00"W
Bearing: 60 deg.

Display Position:
Lat: 32°36'20"N
Lon: 116°54'32"W
Bearing: 60°00'00"
Range: 73.41 nml
Height: 3,125.00 ft
Loss: 130.5 dB

FIGURE 6. AREPS communications application.



Classification
● None
○ Level 1
○ Level 2
○ Level 3

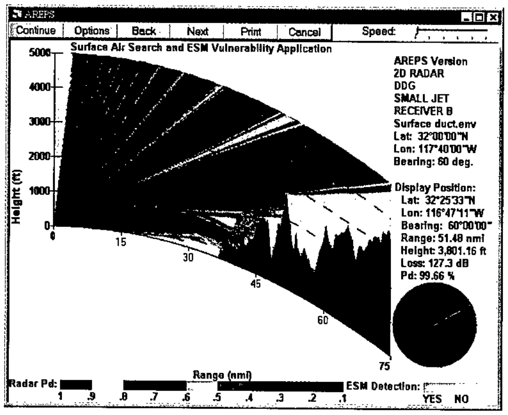
Radar Calculations/Type
● Simple pulsed radar
○ Integrated - Incoherent
○ Integrated - Coherent
○ CW or other radar

Antenna Parameters

Your identification label
Frequency (MHz)
Peak power (kW)
Pulse length (us)
Compressed pulse length (us)
Receiver noise figure (dB)
Assumed system loss (dB)
Max instrumented range (nml)
Pulse rate (Hz)
Free-space range (nml)
1.0E-04 Probability of false alarm

OMNI Antenna type
● Horz ○ Vert ○ Circ Polarization
Hits per scan
Antenna gain (dBi)
Antenna scan rate (rpm)
Horizontal beam width (deg)
Vertical beam width (deg)
Antenna elevation angle (deg)

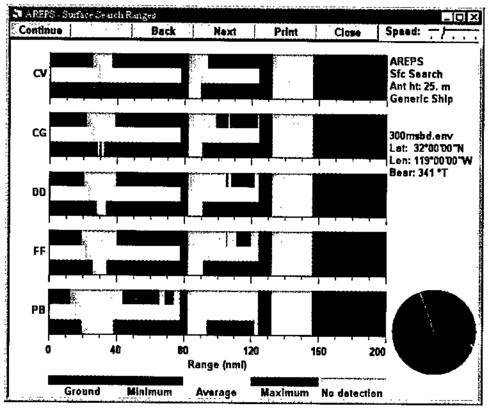
FIGURE 8. AREPS radar system input window.



AREPS Version
2D RADAR
DDG
SMALL JET
RECEIVER B
Surface duct.env
Lat: 32°00'00"N
Lon: 117°40'00"W
Bearing: 60 deg.

Display Position:
Lat: 32°25'33"N
Lon: 116°47'11"W
Bearing: 60°00'00"
Range: 51.48 nml
Height: 3,801.16 ft
Loss: 127.3 dB
Pd: 99.66 %

FIGURE 5. AREPS radar probability of detection and ESM vulnerability application.



AREPS
Sfc Search
Ant ht: 25. m
Generic Ship

300msbd.env
Lat: 32°00'00"N
Lon: 119°00'00"W
Bear: 341 °T

FIGURE 7. Surface-search range tables.

for a radar system. As one navigates the form, input prompts, parameter limits, and other guidance are displayed in a status bar located at the bottom of the window.

AREPS capabilities include antenna radiation patterns of specific system height-finder antennas and a user-defined antenna pattern. Detection threshold calculations include radars using incoherent and coherent integration techniques.

In addition to pulsed radar systems, users may enter continuous wave and other non-pulsed systems, UHF and VHF communications systems, ESM receivers, and radar target descriptions.

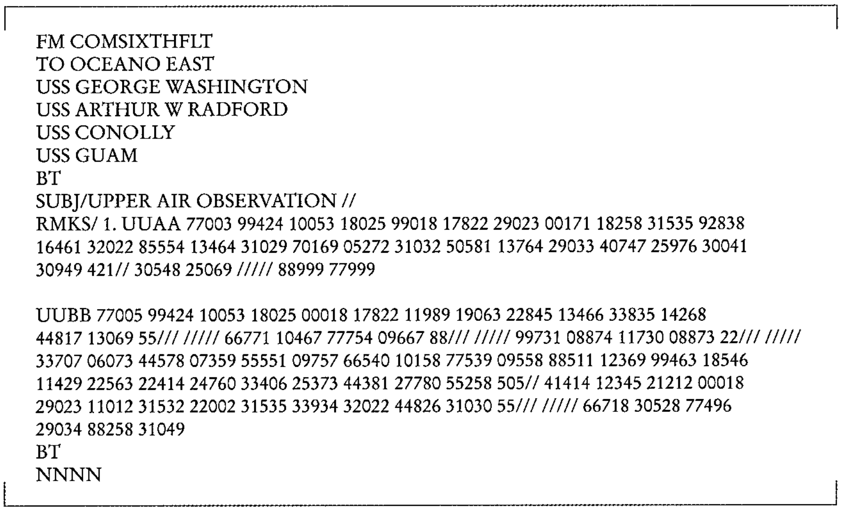
Terrain Data

AREPS derives its terrain height data primarily from the Digital Terrain Elevation Data (DTED) provided by the National Imagery and Mapping Agency (NIMA), available either on CD-ROM or from the NIMA Internet homepage. DTED data are provided in level 0, level 1, and level 2 formats. Level 0 data spacing is 30 arc seconds in horizontal resolution (approximately 1 km). DTED level 0 data are unlimited distribution and may be obtained directly from NIMA's Internet homepage. DTED level 1 data spacing is 3 arc seconds in horizontal resolution (approximately 100 m). Level 2 data spacing is 1 arc second in horizontal resolution (approximately 30 m). Level 1 and 2 data are limited distribution. DTED data are not and may not be distributed with AREPS. For ease of input when using DTED CD-ROMs, users need only specify the latitude and longitude location of their transmitter. The AREPS program will determine which CD-ROM is required, prompt to insert the CD-ROM into the drive, and automatically extract the terrain data needed.

In addition to terrain elevations, the APM allows for the specification of range-dependent surface conditions should users be concerned about surface types for vertically polarized antennas. AREPS uses the surface conditions as defined by the International Telecommunication Union, International Radio Consultative Committee (CCIR). These conditions are provided by plain-language descriptors, selected from a drop-down menu.

Environmental Input

Atmospheric data may be derived from World Meteorological Organization (WMO) upper air observations. The entry of environmental data into AREPS has been completely automated by using the capabilities of the Windows 95/NT operating system. Within normal naval message traffic, WMO-coded radiosonde messages are routinely available. Figure 9 shows such a message.



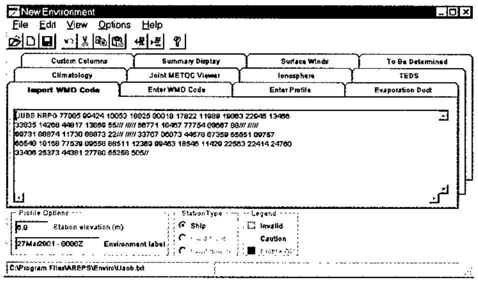
```
FM COMSIXTHFLT
TO OCEANO EAST
USS GEORGE WASHINGTON
USS ARTHUR W RADFORD
USS CONOLLY
USS GUAM
BT
SUBJ/UPPER AIR OBSERVATION //
RMKS/ 1. UUAA 77003 99424 10053 18025 99018 17822 29023 00171 18258 31535 92838
16461 32022 85554 13464 31029 70169 05272 31032 50581 13764 29033 40747 25976 30041
30949 421// 30548 25069 ///// 88999 77999

UUBB 77005 99424 10053 18025 00018 17822 11989 19063 22845 13466 33835 14268
44817 13069 55/// ///// 66771 10467 77754 09667 88/// ///// 99731 08874 11730 08873 22/// /////
33707 06073 44578 07359 55551 09757 66540 10158 77539 09558 88511 12369 99463 18546
11429 22563 22414 24760 33406 25373 44381 27780 55258 505// 41414 12345 21212 00018
29023 11012 31532 22002 31535 33934 32022 44826 31030 55/// ///// 66718 30528 77496
29034 88258 31049
BT
NNNN
```

FIGURE 9. WMO radiosonde message from Commander, Sixth Fleet.

Users need only locate the message (for a ship, the message is usually available on the ship's local area network); open the message file using any ASCII text editor (e.g., Notepad) provided with Windows 95/NT; "copy" the text to the Windows clipboard; and "paste" it into the Import WMO Code window of AREPS (Figure 10).

All extraneous text is filtered; the message is decoded; and a height vs. M-unit profile is automatically created. Should the observation be from a sea-based platform, the surface temperature and humidity are used to calculate a neutral-profile evaporation duct profile, and this profile is appended to the upper air portion of the observation. If surface observations are available, users may override the neutral profile and include full stability dependency.



```
UUBB NRPG 77005 99424 10053 18025 00018 17822 11989 19083 22945 13466
33835 14208 44817 13069 55/// ///// 88771 10467 77754 09687 88/// /////
99731 08874 11730 08873 22/// ///// 33707 06073 44678 07359 65651 09767
66540 10158 77539 09558 88511 12389 99463 18546 11429 22563 22414 24760
33406 25373 44381 27780 55258 505//
```

0.0 Station elevation (m)

27Mar2001 - 0000Z Environment label

● Ship

C:\Program Files\AREPS\Enviro\Uaob.txt

FIGURE 10. Import WMO code tab for new environmental input.

It is not always necessary to have access to a local area network for the WMO observation. Many shore organizations and ships post their local radiosonde observations on their Internet or SIPRNET (Secure Internet Protocol Router Network) homepage. Once such a homepage is found for the user's particular area of interest, the WMO report may be copied to the Windows 95/NT clipboard directly from the browser (such as Netscape or Microsoft Internet Explorer), and then pasted into the Import WMO Code window. For military users, WMO reports are also available from the Fleet Numerical Meteorology and Oceanography Center (METOC) by using the Joint METOC Viewer (JMV) and/or the METOC Broadcast (METCAST) client.

For those without access to observational data in the WMO format, AREPS contains options to import observational data in a generic column format. Should real-time data be unavailable, AREPS contains a climatology of ducting conditions taken from 921 observing stations worldwide.

With the release of AREPS version 3.0, environmental data may now be obtained from mesoscale numerical meteorological models such as the Coupled Ocean and Atmosphere Mesoscale Prediction System (COAMPS). Thus, for the first time, predictions of systems' performance based on future atmospheric conditions are possible, giving the operator or the tactical decision-maker a valuable tool for mission planning.

Distribution and Support

AREPS is configured for Defense Information Infrastructure Common Operating Environment (DII COE) compliance and has been submitted as a Global Command and Control System–Maritime (GCCS–M) segment. We also provide distribution and technical support for the AREPS program. Distribution is provided on CD-ROM through U.S. mail or by direct download of the program from our Internet homepage (http://sunspot.spawar.navy.mil). In addition to the program software, our homepage includes help topics, frequently asked questions, and program service packages.

❖

Wayne L. Patterson

MS in Meteorology, Naval Postgraduate School, 1977

Current Research: Implementation of advanced atmospheric electromagnetic wave propagation models into fleet operational systems; mesoscale meteorological model interfaces to propagation models.

A Passive Ranging Technique for Objects within the Marine Surface Layer

Stephen Doss-Hammel
SSC San Diego

ABSTRACT

Infrared Search and Track (IRST) systems are important to the surface Navy for the detection of low-flying missile threats. Infrared signals propagating within the marine atmospheric surface layer are frequently distorted by strong vertical fluxes. One particular distortion that occurs commonly is the sub-refractive mirage. During sub-refractive mirage conditions, an imaging sensor or camera will record two distinct images of a single point source. A sub-refractive mirage image can be exploited to provide both height and range information. A technique for passive ranging is described, and a case study using field test data is presented as an example of the concept.

INTRODUCTION

Infrared Search and Track (IRST) systems are designed to operate within the marine atmospheric surface layer. This environment can be difficult for radar systems. A reliable passive infrared (IR) system has the potential to provide useful target detection data.

However, the near sea surface environment can also distort images in the infrared. In particular, refraction effects have a strong effect on IR systems, and the occurrence of mirages is not uncommon. This report describes work to exploit one type of mirage, the inferior mirage, to determine range and height of the source creating the mirage image.

REFRACTIVE EFFECTS AND RAY-TRACE TECHNIQUES

The primary computational tool chosen for the analysis of refractive effects was a widget-based simulator called IRWarp that predicts refractive effects [1]. IRWarp uses meteorological conditions as input data for a ray-trace module [2]. The ray-trace data are used to generate detailed information about geometrical transformations induced by the propagation environment.

The ray-tracing method used within IRWarp is from a model by Lehn [3]. The radius of curvature r of a ray is given by:

$$r = \frac{nT^2}{\alpha(\lambda)(T\rho g + p\,dT/dz)} \qquad (1)$$

where T = absolute temperature, ρ = density, p = pressure, g = gravitational acceleration, n = refractive index, and $\alpha(\lambda) = (77.6 + 0.584/\lambda^2) \times 10^{-6}$ for wavelength = λ. It is also assumed that the ray slope does not exceed 10 milliradians.

The formulation in Eq. (1) applies to visible and infrared wavelengths. Pressure A is relatively constant for the measurements made, and the prime determinant of the radius of curvature of near-horizontal rays was the vertical temperature gradient. The ray-trace algorithm first defines the vertical temperature profile as a set of discrete layers, each with a characteristic temperature gradient and refractivity gradient. A characteristic radius of curvature is then assigned to each layer using Eq. (1).

The vertical temperature profile is based upon a surface-layer similarity theory developed by Monin and Obukhov. For the current study, an approach was followed based upon bulk methods for calculating turbulence parameters described by Davidson et al [4]. Field measurements were taken at the sea surface, and at a reference height, and these values were used to determine the particular values of the scaling parameters. Thus, the sea surface temperature would be T_0, and the temperature $T(z)$ at a height z above the water surface would be given by

$$T = T_0 + T_*\left[\frac{\ln(z/Z_{0T} - \psi_T(z/L)}{\alpha_T k}\right]$$

where Z_{0T} is the roughness length for the temperature profile, T_* is the potential temperature scaling parameter, and α_T is the ratio of heat transfer to momentum transfer at the surface. L is the Monin–Obukhov length, and $\psi_T(z/L)$ is a stability correction function.

A ray trace can be generated from the temperature profile by determining a characteristic radius of curvature for each horizontal layer using Eq. (1). Figure 1 displays the traced rays from the ray-trace algorithm for a coordinate system transformed so that the sea surface is the flat x-axis. The figure shows a ray-trace generated from field test temperature profile data. The air–sea temperature difference was ≈−3.5 K. The number of rays has been reduced to make the graphic more legible. The apparent kinks in some of the more sharply bent rays are an artifact; the actual path for the ray is a carefully determined smooth curve, but points on the path are saved only intermittently as needed for the calculation.

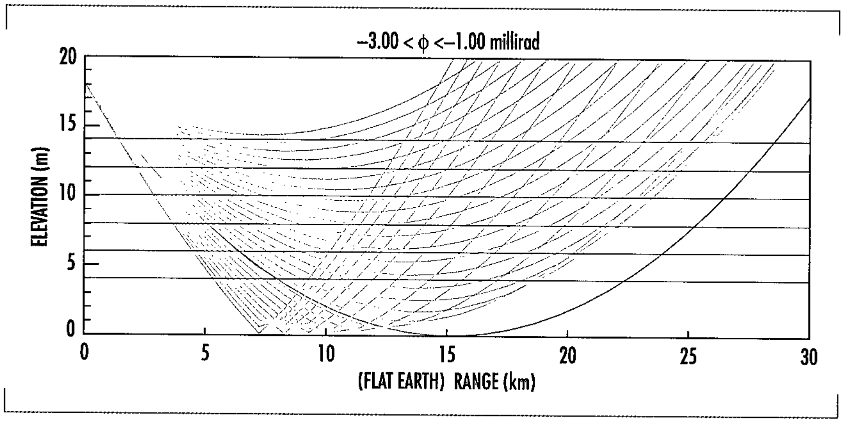


FIGURE 1. The vertical scale is in meters above the surface, while the horizontal scale is in kilometers downrange from the sensor (height = 18 m). The six red horizontal lines indicate level surfaces. The rays that appear to be reflecting from the x-axis are actually refracted.

An atmospheric surface layer for which the air–sea temperature difference is negative exhibits a crucial feature: the rays form a local coordinate system starting at some point downrange. The logarithmic temperature profile ensures that lower elevation rays are deflected to intersect upper elevation rays. The existence of a locally non-degenerate coordinate system implies that in some region of range-height space there exists a one-to-one correspondence with an upper elevation–lower elevation pair that is unique to that point.

TRANSFORMING IMAGE ELEVATION TO HEIGHT-RANGE DATA

The set of rays tracing the propagation path defines an envelope. The ray envelope has an intersection structure with a set of constant-height surfaces (see Figure 1) at heights of 4, 6, 8, 10, 12, and 14 m. A ray traced from the receiver intersects a given constant-height surface either once, twice, or not at all. The intersection structure of the constant-height surfaces with the ray-trace envelope induces a transformation.

To understand transformation more completely, consider Figure 2. The term "isomet" (isomet surface ≡ surface of constant height) is used to refer to the contour curves representing the intersection set between a constant-height surface and the ray-trace envelope shown in Figure 2. Each of the isomets in Figure 2 displays a similar form. The vertical axis shows angular displacement from the horizontal tangent plane at the sensor. The horizontal axis shows range.

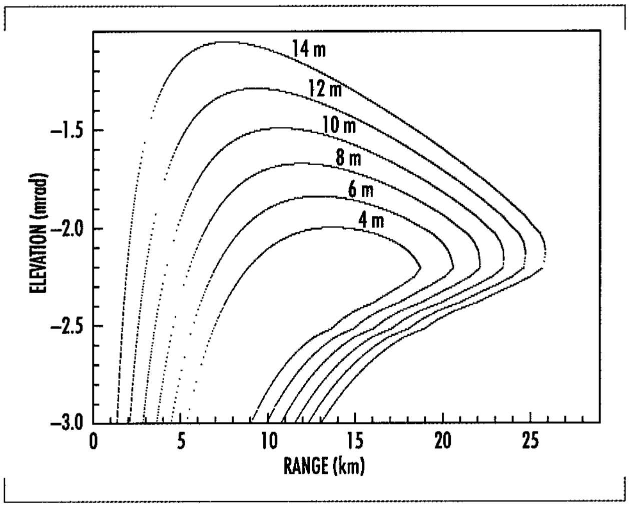


FIGURE 2. A series of isomets at the heights of 4, 6, 8, 10, 12, and 14 m. For a given range value, each isomet defines either 0, 1, or 2 corresponding elevation values.

The graph of a single isomet can be interpreted by imagining a source confined to one of the isomet surfaces (for example, the 14-m isomet) and moving toward the sensor from the 30-km range. At ≈26 km, the source appears over the horizon as a single point that immediately splits into two images. As seen through an imaging sensor, for example, one image decreases in angular elevation, and the upper image increases in angular elevation as the source moves closer in range. At ≈13 km, the lower image descends below –3 milliradians; in terms of the imaginary sensor, it has descended beneath the lowest edge of the sensor focal plane. The (now solitary) upper image continues to rise to the upper edge of the sensor field of view. Within the last 6 km, the source is seen to rapidly move from near the top edge to disappear below the bottom edge.

This form for the 14-m isomet is characteristic of all the isomet contours for surfaces of height less than the sensor height. When the isomet surface height is greater than sensor height, an inbound upper image disappears across the upper boundary, and never re-crosses from top to bottom.

The key to a deduction of height and range from angular elevation information is the utilization of those portions of an isomet for which two values of elevation correspond to a single range value. Thus, for the 14-m isomet, ranges between 13 km and 25 km correspond to two distinct elevation values. This indicates that it is possible to find a one-to-one correspondence between a pair of elevation angles, and a height-range pair.

Thus, the central result in this paper is the transformation shown in Figure 3. When a sensor detects two images, the elevations of the lower and upper images can be plotted as a point in Figure 3, and the height and range of that point can be read from the inner coordinate system. To say it differently, the figure contains the transformation that takes two elevation measurements as input, and generates as output both height and range of the source or target. In terms of coordinate systems, the rectilinear lower elevation vs. upper elevation coordinate system is transformed to the distorted, curvilinear height vs. range coordinate system.

Consider as an example an imaging sensor system with a telescope that detects a source in a

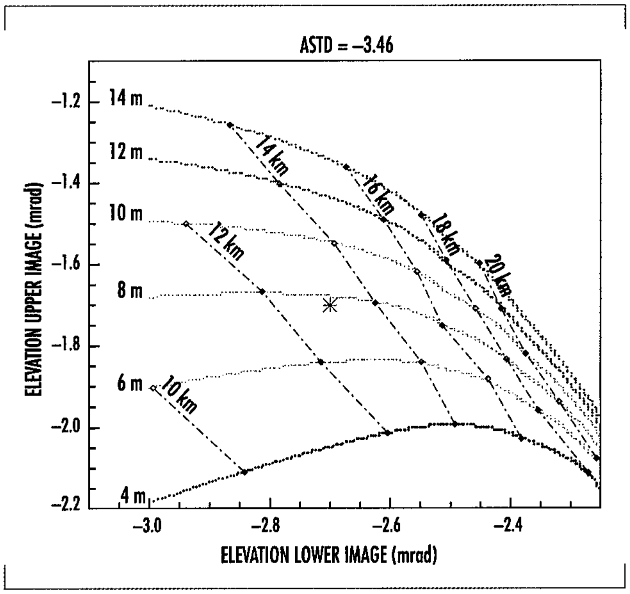


FIGURE 3. The transformation that is implied by the data in Figure 2. The same information is shown here, but restricted to the portions of the isomets that are dual-valued. The point $(\theta_{\text{lower}}, \theta_{\text{upper}}) = (-2.7, -1.7)$ is plotted as an example, and it transforms to range ≈ 13 km, height ≈ 7.5 m.

sub-refractive mirage regime. The two elevations can be determined from the imaging frame: suppose (θ_{lower},θ_{upper}) = (–2.7,–1.7). This example is plotted in Figure 3. Using the transformation, the actual range and height can be read out from the transformed coordinate system, yielding range ≈13 km and height ≈7.5 m.

SUMMARY

Sub-refractive conditions are quite common for the marine atmospheric surface layer. These conditions cause mirages that appear at two different elevations. These two elevations can be transformed by means of a ray-trace technique to yield height and range information.

The usable range for the particular example presented here is from 10 or 12 km out to ≈20 km. Note that the range limits for effective range-finding are determined by the intensity of the sub-refractive conditions. As air–sea temperature difference $T_{air} - T_{sea}$ becomes more negative, the range domain for which two images occur increases in extent by moving the point of first appearance of two images closer to the sensor. Conversely, as air–sea temperature difference $T_{air} - T_{sea}$ becomes less negative and closer to zero, the range domain for which two images occur decreases in extent; the first appearance of two images occurs at a point farther away from the sensor.

Numerous issues remain to be explored. It is necessary to define the limits of applicability for the method. It is also necessary to establish a mathematical foundation for assumptions made concerning the behavior of surfaces and the intersections between them. Furthermore, the method uses an implicit assumption of homogeneity: the full propagation range is characterized by one vertical profile. This appears to be a reasonable assumption for the sub-refractive case, but this also must be carefully examined. To make practical use of the passive ranging technique, it is important to calculate the limits to the angular elevation resolution.

REFERENCES

1. Hammel, S. and N. Platt. 1995. "Topological Description of Mirage Effects," *Proceedings of the Vision Geometry Conference*, SPIE, vol. 2573, pp. 398–406.
2. Platt, N., S. Hammel, J. Trahan, H. Rivera. 1996. "Mirages in the Marine Boundary Layer—Comparison of Experiment with Model," *Proceedings: IRIS Passive Sensors*, vol. 2, pp. 195–210.
3. Lehn, W. 1985. "A Simple Parabolic Model for the Optics of the Atmospheric Surface Layer," *Applied Math. Model*, vol. 9, p. 447.
4. Davidson, K., G. Schacher, C. Fairall, and A. Goroch. 1981. "Verification of the Bulk Method for Calculating Overwater Optical Turbulence," *Applied Optics*, vol. 20, p. 2919.

❖

Stephen Doss-Hammel

Ph.D. in Applied Mathematics, University of Arizona, 1986

Current Research: Atmospheric optics and dynamical systems.

Silicon-on-Sapphire Technology: A Competitive Alternative for RF Systems

Isaac Lagnado and Paul R. de la Houssaye
SSC San Diego

S. J. Koester, R. Hammond, J. O. Chu, J. A. Ott, P. M. Mooney, L. Perraud, and K. A. Jenkins
IBM Research Division, T. J. Watson Research Center

ABSTRACT

We investigated the formation of high-performance, device-quality, thin-film silicon (30 to 50 nm) on sapphire (TFSOS) for application to millimeter-wave communication and sensors. The resulting TFSOS, obtained by Solid Phase Epitaxy (SPE), and the growth of strained silicon-germanium (SiGe) layers on these TFSOS demonstrated enhanced devices and, hence, integrated-circuit performance not achieved previously. We fabricated 250-nm and 100-nm T-gated devices with noise figures as low as 0.9 dB at 2 GHz and 2.5 dB at 20 GHz, with G_a of 21 dB and 7.5 dB, respectively. 250-nm devices resulted in distributed wideband amplifiers (10-GHz bandwidth [BW], world record) and tuned amplifiers (15-dB, 4-GHz BW). 100-nm devices produced voltage controlled oscillators (VCOs) (25.9-GHz), 30-GHz frequency dividers. We obtained f_t (f_{max}) of 105 GHz (50 GHz) for n-channel and 49 GHz (116 GHz, world record) for p-MODFETs (strained $Si_{0.2}Ge_{0.8}$ on a relaxed $Si_{0.7}Ge_{0.3}$ hetero-structure). This paper details our investigation and provides cost comparisons with competing technologies.

INTRODUCTION

Device-quality, thin-film silicon-on-sapphire (TFSOS), obtained by Solid Phase Epitaxy (SPE), has achieved truly outstanding results that are incorporated into present and future high-performance products, such as phase-locked loop integrated circuits (ICs) for wireless communications [1], single-chip Global Positioning System (GPS) receivers, and analog to digital converters (A/DCs) for space applications.

EXPERIMENTAL STUDY

Table 1 shows the measured performance of a front-end receiver at 2.4 GHz; and voltage-controlled oscillators (VCOs), frequency dividers, and tuned amplifiers at >20 GHz. The n-metal oxide semiconductor (MOS) VCO at 26 GHz [2] has the highest tuned frequency ever achieved among complementary metal-oxide semiconductor (CMOS) VCOs.

Concomitantly, recent advances in SiGe epitaxial growth technology indicate that SiGe-based strained-layer modulation-doped field-effect transistors (MODFETs) may be promising alternatives to III–V metal-semiconductor field-effect transisitors (MESFETs) and high-electron mobility transistors (HEMTs) for future high-speed analog communications applications. Electron and hole mobilities well in excess of bulk Si mobilities can be realized in tensile-strained Si quantum wells (QWs) [3] and compressive-strained SiGe [4] or pure Ge QWs [5], respectively. Note that p-MODFETs have demonstrated dc and RF performance figures comparable to n-MODFETs, suggesting the possibility of very-high-speed complementary operation [6], a capability not available in current III–V technology. In this work, we have applied this knowledge to the growth of SiGe relaxed buffer layers and the fabrication of SiGe strained-layer MODFETs on TFSOS. One of the benefits of using insulating substrates, such as sapphire, is the potential solution to the reduction of high losses seen in the microwave frequency regime due to the conducting nature of the silicon substrate. Here, we demonstrate the development of the epitaxial growth of high-mobility, modulation-doped, composite-channel heterostructures on silicon-on-sapphire (SOS) substrates, and describe the resultant outstanding RF characteristics of ≤100 nm T-gate p-MODFETs fabricated on these layer structures.

The composite-channel heterostructure device has the basic structure shown in Figure 1. This layer structure was grown on an SOS substrate

TABLE 1. Measured performance of various components of a front-end receiver designed to operate near 2.4 and 18 GHz.

Measured	Operating Frequency	Gain	NF (50 Ω)	IP3 (output)	Power@Vdd
LNA	2.4 GHz	11 dB	2.2 dB	14 dBm	13.2 mW@1.5V
LNA (HEMT, 2 stages)	1.4–2.6 GHz	25 dB	2.3 dB	15 dBm	
Mixer	Center = 2.4 GHz IF = 250 MHz	-5 dB		5 dBm	8.4 mW @1.5V
Mixer (HBT)	1.4–2.6 GHz	-4 dB	15 dB	0 dBm	
VCO	25.9 GHz 0.6-GHz tuning range		-106 dBc/Hz (phase noise)		24 mW @1.5V
Frequency Divider	1.5–20 GHz 5.9–26.5 GHz				29.5 mW (Core) <20 mW (Core)
Tuned Amplifier	23 GHz 4-GHz Bandwidth	6–7 dB			

as well as a bulk Si control wafer. The devices had a gate length, L_g, of 100 nm.

The room-temperature Hall mobility and sheet carrier density [7] of the composite-channel layer structure grown on an SOS wafer were 800 to 1200 cm^2/Vs and 3.1-2.5 x 10^{12} cm^{-2} at room temperature, respectively, as shown in Figure 2.

The room-temperature output (transfer) characteristic for 100-nm gate-length devices showed practically no difference between devices fabricated on SOS and Si control devices (Figure 3).

Figure 3 also shows that the best SOS transistor had a maximum extrinsic transconductance of 377 mS/mm, which is, to the authors' knowledge, the highest ever reported for alloy-channel p-MODFETs. The device had a corresponding output conductance of 25 mS/mm leading to a maximum dc voltage gain of 15. The only apparent degradation of the device performance caused by the SOS substrate was roughly an order of magnitude higher gate-leakage current compared to the Si monitor, a result that we again attribute to the increased defect density of the SOS wafers.

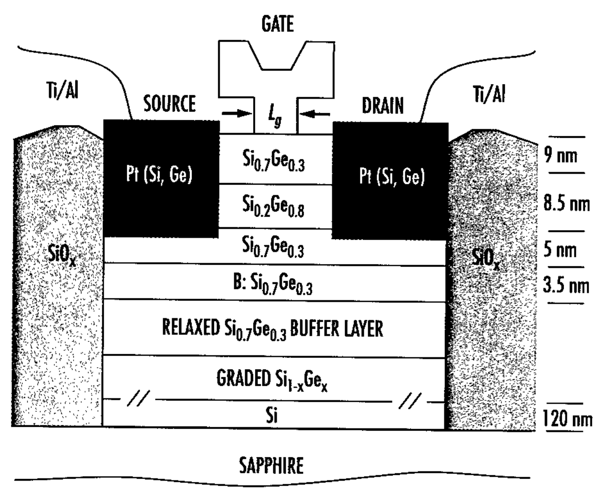


FIGURE 1. Cross-sectional diagram of epitaxial layer structure and p-MODFET device design.

Figure 4 shows frequency-dependent plots [8] of the forward current gain ($|h_{21}|^2$) and the maximum unilateral gain (MUG) for a 0.1 x 50 μm^2 p-MODFET on SOS. Values of f_T = 49 GHz and f_{max} = 116 GHz were obtained after de-embedding the contact pads; the latter value being the highest f_{max} ever reported for a SiGe p-MODFET. Figures 5 and 6 illustrate the fact that f_t and f_{max} saturate at a very low bias voltage; f_{max} reached 100 GHz at V_{ds}~0.6 V and remained sustained over a wide bias range. In Figure 7, the small-signal parametric model reveals non-negligible capacitances (C_{pg} and C_{pd}) caused by the incomplete removal of the SiGe buffer layer and Si film in the isolation regions. Through the agreement of the raw data and extracted values, Figure 8 demonstrates the accuracy

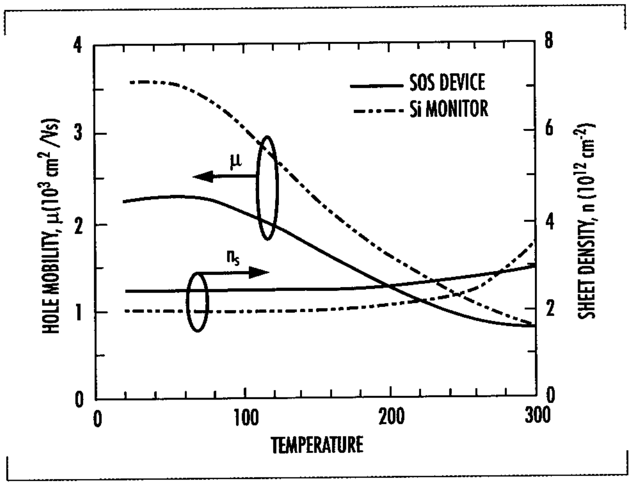


FIGURE 2. Hole mobility and sheet density *vs.* temperature for composite-channel layer structures grown on SOS and Si control wafers.

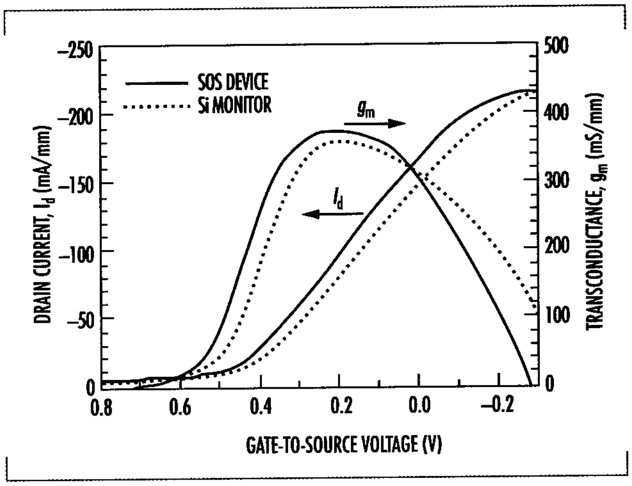


FIGURE 3. Comparison of 0.1-μm composite-channel p-MODFETs on Si and SOS. The bias voltage is V_{ds} = –0.6 V.

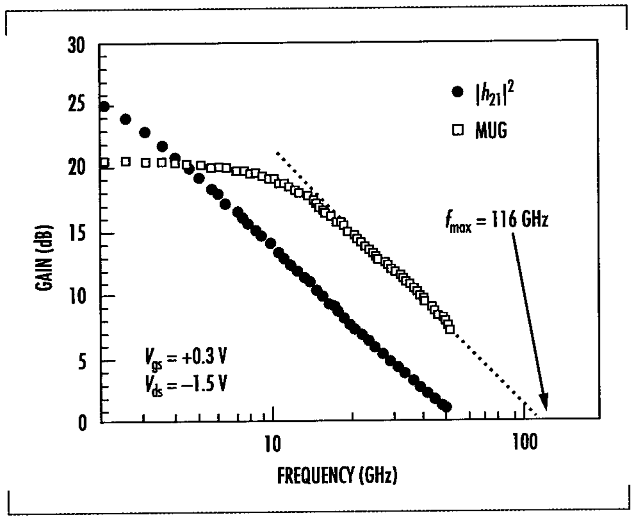


FIGURE 4. Plot of $|h_{21}|^2$ and MUG *vs.* frequency for a 0.1 x 50 μm^2 composite-channel p-MODFET on SOS. Values of f_t = 49 GHz and f_{max} = 116 GHz are obtained after open-circuit de-embedding.

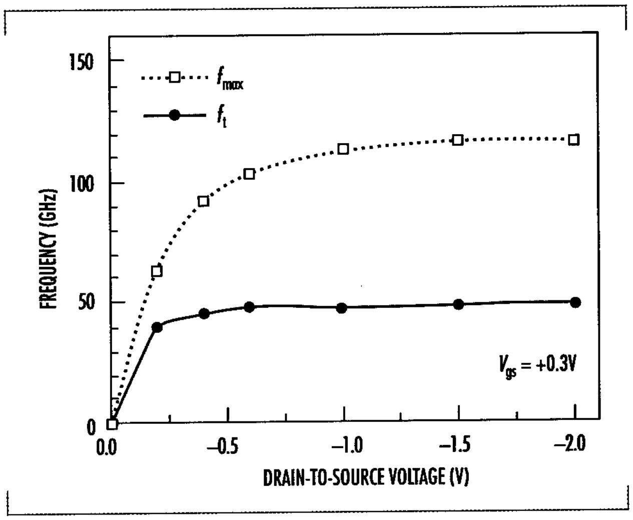


FIGURE 5. Bias dependence of f_t and f_{max}. f_t and f_{max} saturate at a very-low-bias voltage.

of the premliminary device model created from these results. Design of circuits from this model should demonstrate the unprecedented potential of SiGe/TFSOS technology.

Table 2 lists the comparative cost of competing technologies [9, 10, and 11] for the manufacturer and the user.

CONCLUSION

The incorporation of strained SiGe heterostructures on thin-film silicon-on-sapphire (the device-quality Si film obtained either through SPE or layer bonding) for n- and p-FETs, characterized by superior transport carrier properties, high dynamic performances (f_t, f_{max}), and low noise at high frequencies (Figure 9) will enable an entirely new technology. The

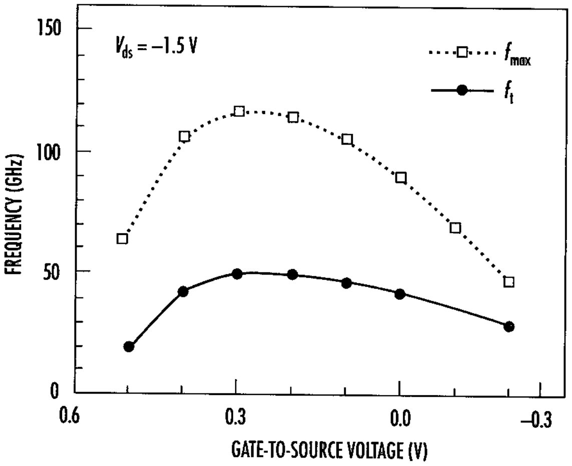


FIGURE 6. Bias dependence of f_t and f_{max}. High f_{max} sustained over wide bias range.

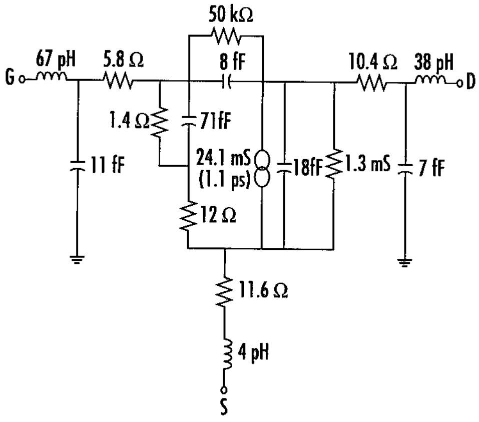


FIGURE 7. Small-signal equivalent circuit. Small-signal parameters reveal non-negligible capacitances (C_{pg} and C_{pd}) from unremoved SiGe buffer layer in isolation regions.

TABLE 2. Relative cost of different technologies as seen by (A) the manufacturer and (B) the user. The difference is due to the different profit margins available to the companies as determined by what the market will bear.

A. Manufacturing Cost [9, 10]

Si Bulk CMOS	SOS CMOS	Bipolar Si & SOS SiGe	GaAs MESFET	HBT(GaAs)
1	1.3	3.5	3.5–7	10

B. User Cost [11]

Technology	Cost per sq. mm ($US)
Silicon CMOS	0.01
SiGe epitaxy	0.60
GaAs epitaxy	2.00
InP epitaxy	10.00
Tokyo real estate	0.01

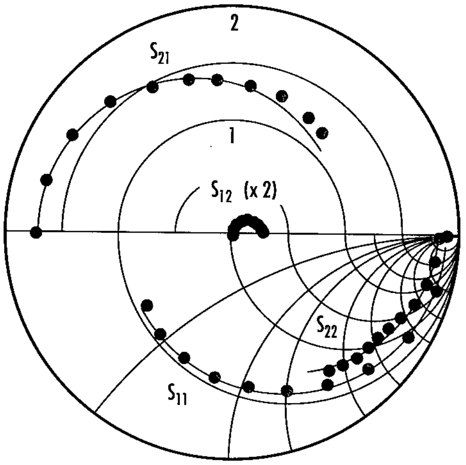


FIGURE 8. Comparison of s-parameter. Good agreement between raw data (points) and extracted values (lines).

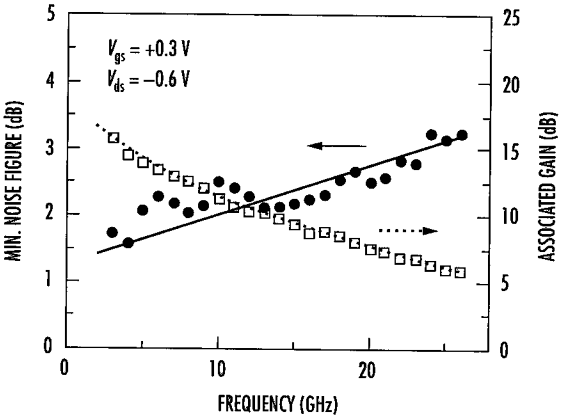


FIGURE 9. Noise parameter results. Values of F_{min} = 2.5 dB and G_a = 7.5 dB obtained at 20 GHz. Low-frequency noise dominated by gate-leakage current.

resulting impact of the combined TFSOS and SiGe technology on the marketplace, both nationally and internationally, will be quite revolutionary because no other material can provide a complementary technology as efficiently, from either the technical or economic aspect.

REFERENCES

1. Peregrine Semiconductor Corporation data sheet.
2. Wetzel, M., L. Shi, K. Jenkins, P. R. de la Houssaye, Y. Taur, P. M. Asbeck, I. Lagnado. 2000. "A 26.5-GHz Silicon MOSFET 2:1 Dynamic Frequency Divider," *IEEE Microwave and Guided Wave Letters*, vol. 10, no. 10, pp. 421–423.
3. Nelson, S. F., K. Ismail, J. O. Chu, and B. S. Meyerson. 1993. "Room-Temperature Electron Mobility in Strained Si/SiGe Heterostructures," *Applied Physics Letters*, vol. 63, pp. 367–369.
4. Ismail, K., J. O. Chu, and B. S. Meyerson. 1994. "High Hole Mobility in SiGe Alloys for Device Applications," *Applied Physics Letters*, vol. 64, pp. 3124–3126.
5. Höck, G., M. Glück, T. Hackbarth, H.-J. Herzog, and E. Kohn. 1998. "Carrier Mobilities in Modulation Doped Si_{1-x} Ge_x Heterostructures with Respect to FET Applications," *Thin Solid Films*, vol. 336, pp. 141–144.
6. Armstrong, M. A., D. A. Antoniadis, A. Sadek, K. Ismail, and F. Stem. 1995. "Design of Si/SiGe Heterojunction Complementary Metal-Oxide-Semiconductor Transisitors," *International Electron Devices Meeting (IEDM) Technical Digest*, pp. 761–764.
7. Koester, S. J., R. Hammond, J. O. Chu, J. A. Ott, P. M. Mooney, L. Perraud, and K. A. Jenkins, I. Lagnado, P. R. de Ia Houssaye. 1999. "High-Performance SiGe pMODFETs Grown by UHV-CVD," *Proceedings of the 7th International Symposium on Electron Devices for Microwave and Optoelectronic Applications (EDMO 99)*, 22 to 23 November, King's College, London, England.
8. Koester, S. J., R. Hammond, J. O. Chu, P. M. Mooney, J. A. Ott, C. S. Webster, I. Lagnado, and P. R. de la Houssaye. 2000. "Low-Noise SiGe pMODFETs on Sapphire with 116-GHz f_{max}," *58th IEEE Device Research Conference*, 19 to 21 June, Denver, CO.
9. Young, J. P. and R. Collins. 1997. Course given at *IEEE Microwave Theory and Techniques Symposium (MTT-S).*
10. Wienau, D. 2000. "SiGe RF Technology for Mobile Communications—Technical and Commercial Aspects," *Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems: Digest of Papers*, pp. 79–82.
11. Paul, J. D. 1999. "Silicon-Germanium Strained-Layer Materials in Microelectronics," *Advanced Materials*, vol. 11, no. 3, pp. 191–204.

❖

Isaac Lagnado

Ph.D. in Solid-State Electronics, MIT, 1970

Current Research: Thin-film silicon-on-sapphire (TFSOS); silicon germanium (SiGe) on TFSOS; application to RF mixed-signal systems on a single chip (SOC); high performance analog-to-digital converters; high-efficiency, high-power density power supplies.

Paul R. de la Houssaye

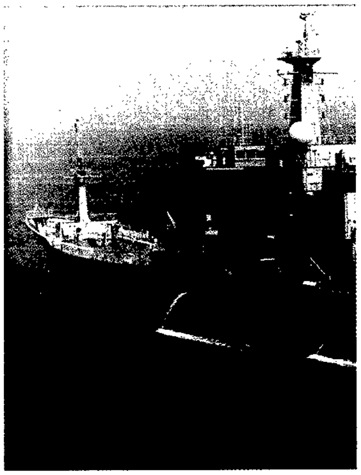
Ph.D. in Applied Physics, Stanford University, 1988

Current Research: Thin-film silicon/SiGe on sapphire device and circuit fabrication; novel microelectromechanical systems (MEMS).

6

Navigation and Applied Sciences

■



6 NAVIGATION AND APPLIED SCIENCES

An Integrated Approach to Electronic Navigation

Peter Shaw and Bill Pettus
SSC San Diego

ABSTRACT

While the Global Positioning System (GPS) is and will continue to be an excellent navigation system, it is neither flawless nor is it the only system employed in the navigation of today's seagoing warfighters. The modern warfighter must operate with dominant maneuverability, precision engagement capability, full-dimensional protection, and focused logistics. To meet these requirements, an integration of independent, self-contained, self-initiated, and externally referenced systems must be realized. The Navigation Sensor System Interface (NAVSSI) AN/SSN-6 (V) provides this capability through the real-time collection, processing, and distribution of accurate and reliable positioning, navigation, and timing (PNT) data from varied shipboard sensors and systems. NAVSSI adds to this an electronic navigation capability that provides the ship navigation team with route planning, route monitoring, and safety of navigation capabilities.

INTRODUCTION

The Navigation Sensor System Interface (NAVSSI) integrates inputs from various shipboard navigation sensor systems, distributes the integrated navigation solution to shipboard users, and provides a dedicated workstation to the ship's navigator. NAVSSI uses an open systems architecture, government off-the-shelf (GOTS) software, and commercial off-the-shelf (COTS) hardware [1].

NAVSSI is an Evolutionary Acquisition (EA) program that is entering its fourth phase: the development of Block 3 hardware and software. The new Block 3 configuration is being developed to expand the number of sensor and user systems supported. Block 3 also incorporates a Global Positioning System (GPS) Joint Program Office approved embedded GPS receiver directly into NAVSSI, refines its integrated navigation solution algorithms, and further expands the navigation tools available to the ship's navigation team.

The NAVSSI system is actually an integration of subsystems. The Real-Time Subsystem (RTS), which collects, processes, and distributes the positioning, navigation, and timing (PNT) data, uses a set of navigation source integration algorithms to blend input data from sensors such as GPS and Inertial Navigation Systems (INS) to produce a highly accurate and robust navigation solution. When required, this solution is referenced to the ownship's reference point (OSRP).

The Display Control Subsystem (DCS) provides the operator interface to the RTS. It also contains the electronic charting and navigation capabilities as well as a radar interface and chart product distribution capability. The charting software used is the U.S. Coast Guard developed Command Display and Control Integrated Navigation System (COMDAC INS). The DCS and COMDAC INS software packages are built on the Defense Information Infrastructure Common Operating Environment (DII COE) and together will enable NAVSSI to lead the way to Electronic Chart Display Information System–Navy (ECDIS–N) compliance.

The DCS can display ownship's navigation sensor information, log navigation fix data, use the navigation toolkit, display digital nautical charts and radar contacts, and control the RTS(s). The RTS accepts and integrates data from both external navigation sensor systems and the embedded GPS receiver cards. The RTS distributes real-time navigation data and precise time to shipboard user systems and communication systems.

The DCS communicates with each RTS via a local-area network (LAN). Depending on the particular installation this may be an independent subsystem LAN or an existing shipboard LAN. Many, but not all, installations of NAVSSI Block 3 will include two RTSs. On a dual RTS installation, two RTSs will exchange data via a reflective memory link.

Other system components are the Bridge Workstation (BWS) and NAVSSI Remote Station (NRS), which are the remote displays for the DCS operator. The NAVSSI BWS and NRS have full control and display capabilities to ship's force on the bridge.

The program manager for the development of NAVSSI is the Program Manager for Air and Sea Navigation Systems at the Space and Naval Warfare Systems Command (SPAWAR PMA/PMW 187).

The following sections describe the requirements for system performance and the operational characteristics of the NAVSSI Block 3 System.

REAL-TIME SUBSYSTEM (RTS)

The NAVSSI RTS receives navigation data from multiple sensors and systems and provides real-time output of an integrated navigation solution to multiple user systems. The NAVSSI RTS(s) accepts and processes data in a variety of formats as listed in Table 1.

TABLE 1. Summary of Block 3 data inputs.

System	Message Rate (Hz)$^{\text{Note 1}}$	Data Received
Redundant RTS	50$^{\text{Note 2}}$	All Sensor Data
DCS	Variable	All Sensor Data, Control, Config, Lever Arms, etc.
AN/SPS 73	1	Scanned Radar Images/Contacts
AN/WRN-6 IP	1	PVT, almanac, status
AN/WRN-6 ntds	1	PVT, status
AN/WSN-5 Ch A	4.07	PVT, speed$_W$, performance
AN/WSN-5 Ch B	8.14 or 16.28	PVT, attitude, speed$_W$, performance
AN/WSN-7 GPS/010	4	PVT, attitude, speed$_W$, performance
AN/WSN-7 Superchannel	50$^{\text{Note 2}}$	PVT, attitude, rate speed$_W$, performance
BFTT	1	Training Data
Speed Log$_{\text{digital}}$	8	Speed$_W$
DMS/FODMS	10$^{\text{Note 2}}$	INS, fathometer, wind, propulsion
DSVL	8	Speed$_W$ or speed$_g$
EM Log	Continuous	SpeedW
Fluxgate compass	1	Heading
FOAL receiver	Continuous	RF Input
GVRC	1	PVT, status, almanac
Gyrocompass	Continuous	Synchro heading
ICAN	Wind at 10 Std Msg at 50	Wind speed and direction, std msg
IP-1747/WSN-7	0.125	WSN-7 Control
MK 38 AEGIS Clock Converter	1024	Aegis Combat System Time
MK 39 AEGIS Clock Converter	1024	Aegis Combat System Time
Moriah (NDWMIS)	10	True wind speed and direction
SWAN	50	Std msgs
LPD 17 Wind	10	Wind speed and direction
UQN-4/4A	1	Depth$_{\text{keel}}$

$^{\text{Note 1}}$ Data rates are approximate.

$^{\text{Note 2}}$ Multiple messages, highest rate given.

Data Integrity Checking

NAVSSI continuously monitors the data inputs from each source listed in Table 1 to ensure data integrity. Integrity checking consists of, but is not limited to, ensuring proper reception of data over the physical medium connecting the source to NAVSSI. Incomplete messages, messages with checksum errors, etc., are processed as identified in the appropriate interface design specification (IDS). As required, NAVSSI posts visible alerts to the operator.

Message level validity checking is conducted based on source validity indicators transmitted with the data if the appropriate IDS provides for such indicators. Sources indicating that their data are invalid are not used by NAVSSI until the data are again marked valid by the source.

NAVSSI provides navigation data validity monitoring, consisting of continuous monitoring of the time evolution of each position source's error characteristics. If estimates in the source's error consistently fall outside of the source's statistical performance bounds, the source is marked as invalid. The automatic source integration algorithm does not use these data, and NAVSSI posts a visible alert to the operator. If a NAVSSI operator manually selects a source that is out of its performance bounds, NAVSSI displays a warning message to the operator.

Navigation Source Integration

NAVSSI provides navigation source integration algorithms that blend the input data received from GPS with available INS data to produce a highly accurate and robust navigation solution. The algorithms written to perform navigation source integration take into account the error characteristics associated with each navigation sensor system and will meet the accuracy requirements specified in Table 2. Each RTS resolves navigation position information from each sensor to the same single shipboard reference point. The reference point for this integrated solution is the OSRP. User systems receive data based on the integrated OSRP solution, excepting those systems for which the IDS states that the data shall reflect a specific data source reference point.

The navigation source integration algorithms operate automatically or manually.

Automatic Source Selection Mode

In Automatic Mode, the RTS(s) provides navigation data from the data sources selected by the navigation source integration algorithms to the DCS and external user systems. The Automatic Mode is the default mode.

The navigation source integration algorithms estimate the accuracy of the data being output by NAVSSI. This accuracy estimate is based on the known nominal error characteristics of the available sensor systems, a comparison of the available data and the maintenance of long-term sensor accuracy data. Some user systems are sent the accuracy estimation data as part of their data message. For other user systems, these data are used to determine the setting of validity bits.

The source integration algorithms enable NAVSSI to provide appropriately referenced latitude and longitude accurate to within 12 m (two dimensions, one sigma) under non-casualty conditions. This accuracy requirement is significantly more stringent than the current requirements for INSs and is one of the primary drivers for the redesign of the NAVSSI Block 2 source integration routines into source integration routines in Block 3. The accuracy requirement is based on a root-sum-square of all known error components, including the worst-case latency-related error for each of the interfaces.

NAVSSI integrates velocity input from the various sensors to maintain a real-time estimate of accurate velocity. A maximum ship speed of 40 knots is assumed.

The source integration algorithms enable NAVSSI to provide attitude data from the best available attitude source. On ships that have INSs, NAVSSI tracks the accuracy of the INSs and provides the users with attitude data from the chosen INS. As seen in Table 2, attitude latency is critical for many user systems because the error caused by data latency can quickly exceed the error budget. Therefore, for certain user systems, it is necessary to schedule data output messages to coincide with the receipt of fresh data from the appropriate sensor in order to meet the requirements given in Table 2.

Manual Source Selection Mode

The operator can override the automatic source selection algorithms for the data displayed on the DCS and the data sent to external users. The operator is prompted to choose one or all of the following sources: position data source, velocity data source, attitude data source, or time data source.

When in manual override, the data sent to the DCS, the INS, and the other external users are taken from the manually chosen sources for data. If the operator does not manually choose a source for a particular type of data, those data continue to be provided via the navigation source integration algorithms. In addition, the NAVSSI operator can manually override the INS integration algorithms and choose the best INS.

If there is a loss of communication with a manually chosen source or if the data from that source are marked as invalid, the following are performed:

- the RTS(s) sends an alert message to the DCS operator;
- if position data were manually selected, the RTS(s) estimates position from the last valid message from the manually selected source and provides these data to the DCS and/or users;
- if velocity data were manually selected, the RTS(s) marks the velocity data being sent to the DCS and/or user systems as invalid;
- if attitude data were manually selected, the RTS(s) marks the attitude data being sent to the DCS and/or user systems as invalid;
- if the time source was manually selected, the RTS(s) maintains the last offset calculated from the chosen time source and uses the microprocessor clock to continue to update time.

Position Data Referencing

The lever arms to correct navigation data from each sensor system to OSRP are entered into the RTS(s) system configuration files via the DCS by the installing activity. Once entered, these lever arm data are maintained in non-volatile storage as part of the ship's NAVSSI system configuration files so that the RTS(s) automatically performs OSRP corrections. Thus, all user systems receive position data referenced to OSRP unless the IDS for that system interface specifically designates that a particular sensor's uncorrected data be used for the position data for that system.

Lever arm data from the following systems (if installed) are provided to the RTSs so that the RTSs perform the corrections needed to reference all position data to OSRP: GPS antenna No. 1, GPS antenna No. 2, INS No. 1, and INS No. 2.

GPS reset data sent to the ship's Inertial Navigation Systems are referenced to OSRP. This is done in order to make the system work equally well with AN/WRN-6 or the dual antenna GPS Versa Modular European Receiver Card (GVRC).

If attitude data are not available from the INS, NAVSSI uses the following estimates to complete its OSRP lever arm corrections:

- Heading = TAN-1 (VE/VN) for positive Velocity East (VE) and VN
  = 180° + TAN-1 (VE/VN) for negative Velocity North (VN)
  = 360° + TAN-1 (VE/VN) for negative VE and positive VN
- Roll = 0°
- Pitch = 0°

INS Accuracy Estimation

In most of the Block 3 configurations, NAVSSI communicates with an installed INS. Depending on the installation, this INS can be the Standard Shipboard Inertial Navigation System (AN/WSN-5) or the Ring Laser Gyro Navigator (RLGN or AN/WSN-7).

Each RTS normally communicates directly with only one INS, but will have access to the data from the other INS in dual RTS/INS installations via a reflective memory link with the second RTS. In support of the sensor integration algorithms and as an aid to the ship's navigation team, each RTS independently and continuously evaluates the accuracy of both INSs. This independent assessment of INS accuracy uses GPS data (when available) and enables NAVSSI to estimate INS accuracy both in terms of absolute accuracy and accuracy relative to the other INS. Thus, these routines enable NAVSSI to choose INS data from the more accurate INS. However, the accuracy algorithms will include a minimum 10% allowance for hysteresis. This will prevent NAVSSI from repeatedly switching back and forth between the two INSs when both have relatively similar performance characteristics. These accuracy estimation routines are a significant improvement over the Block 2 INS assessment algorithms, which simply used the accuracy bits provided by each INS and did not attempt to make any independent assessment of INS accuracy.

The RTS(s) can receive and respond to a manual selection of best INS from the DCS.

Estimated Position Processing

The RTS(s) calculates Estimated Position (EP) based on discrete position fixes, best available heading source, and best available speed source. The RTS(s) can also use manually entered course and speed to calculate EP. EP data are provided to the source integration algorithms for consideration as a candidate for source integration and to the DCS for display. However, EP is not used as source data unless manually selected or as the result of multiple sensor failures.

RTS Output Data

The NAVSSI RTS outputs navigation and time data in a variety of formats. Table 2 summarizes the data output requirements and lists interface criticality requirements for maintaining communications in the event of single and multiple point failures.

Output Data Time Tagging

Time information within output data messages is accurate to within 200 msec (two sigma) root-sum-squared with any timing inaccuracy of the sensor data when the output message structure provides sufficient resolution to support this accuracy. In case of a loss of GPS input, NAVSSI is able to maintain time accurate to 1 msec for 1 day and accurate to 10 msec for 14 days.

Precise Time Distribution

The NAVSSI RTS(s) provides accurate time to user systems by means of Have-Quick, Binary Coded Decimal (BCD) time code, Inter-Range Instrumentation Group (IRIG-B) time codes, 1 pulse per second (1 PPS), and 10 pulse per second (10 PPS). The Have-Quick, BCD time code, and 1-PPS signals meet the standards specified in ICD-GPS-060. The IRIG-B time conforms to the standards set forth in IRIG Standard 200-98. The 10-PPS signal, implemented in the NAVSSI Precise Time Unit has all of the characteristics of the 1 PPS signal except that it is at 10 times the rate. Table 2 identifies the accuracy of the time data required by the various user systems. All requirements are in terms of a two-sigma level of accuracy.

TABLE 2. Block 3 data output summary.

System	Message Rate (Hertz)	Position Accuracy (meters)	Attitude Latency (msec)	Time Accuracy (msec)
ACDS Block 0	8	100	0	100
KSQ-1	1	100	N/A	1000
SMQ-11	50	10	20	1000
SPS-73	4 or 1	16	N/A	100
SQS-53d	1	100	N/A	100
SRC-54 Singars	1	N/A	N/A	1
TPX-42	8	10	50	100
WRN-6	4	N/A	100	N/A
WSN-5	1	100	N/A	1
WSN-7	1	100	N/A	1
ATWCS	Multiple	N/A	N/A	N/A
BFTT	1	0	1000	0.1
BGPHES	1	100	N/A	N/A
CADRT	1	N/A	N/A	N/A
CEC	50 [Note 1]	20	10	0.001
CDL-N	8	100	60	100
COBLU	1	100	N/A	N/A
Combat DF	1	100	N/A	N/A
DCS	1	N/A	N/A	N/A
DBB	1	100	N/A	1000
DSVL	8	N/A	250	125
ERGM	1	20	N/A	10

(Table 2 is continued on the following page.)

[Note 1] Multiple messages, highest rate given.

Output NAVSSI Standard Messages

NAVSSI Standard Messages have been created to facilitate future design efforts for use of a wide variety of potential user systems. The Standard Message content is independent of the hardware chosen for any particular interface, allowing this message to be sent at different rates and over a wide variety of point-to-point and LAN interfaces. NAVSSI Navigation Message is a generic sub-message format designed to meet the requirements of various navigation user systems. It includes basic navigation data and time data. Other sub-messages include True Wind, Apparent Wind, Magnetic Variation, Own-Ship Distance, and Navigation Sensor.

Expansion Port Capability

NAVSSI Block 3 has the capability of supporting at least six new users without any software or hardware system modifications. To meet this goal, each Block 3 RTS hardware suite is designed with expansion ports, of which at least two output ports will be Electronic Industries Association (EIA) Standard RS-422 and at least two ports will be MIL-STD-1397 Rev C Type E (Low Level Serial). Electronic cards supporting the expansion ports need not be actually installed, but the internal cabling and backplate connectors do need to be fully prepared. NAVSSI Expansion Ports transmit one of three messages: the accuracy of the data are 25 m for position, 100 msec for time, and 100 msec for attitude latency. For RS-422 interface, the transmission rate is selectable between 1 or 4 Hz. For NTDS-E interface, the transmission rate is selectable as 1, 4, 8, 16, or 50 Hz. The WRN-6 IP message (or Time Mark Data Message), defined in ICD-GPS-150, has a C4 criticality. The accuracy of the data for position is 25 m and 1 sec for time. The data are transmitted at a rate of 1 Hz over RS-422 interface. The NMEA 0183 message, defined in National Marine Electronics Association (NMEA) 0183 Version 2.30, has a C4 criticality. The message, composed of 10 sub-messages, is transmitted at a rate of 1 Hz, with the exception of Heading True (HDT) at 8 Hz. The accuracy of the data is 100 m for position and 1 sec for time.

TABLE 2. Block 3 data output summary (continued).

System	Message Rate (Hertz)	Position Accuracy (meters)	Attitude Latency (msec)	Time Accuracy (msec)
Exp Port	Note 2	25	100	100
FODMS (Stanag)	40	N/A	N/A	N/A
FODMS (RS 422)	1	100	N/A	N/A
ICAN	50$^{\text{Note 1}}$	20	10	10
IP-1747 WSN-7	0.125$^{\text{Note 1}}$	N/A	N/A	N/A
MK-34 MK-160	8$^{\text{Note 1}}$	20	N/A	10
MK-86	1	20	N/A	100
METOC	1	16	100	N/A
Ndwmis	1$^{\text{Note 1}}$	16	100	N/A
NDDN	1	100	N/A	N/A
NMEA	1	100	100	100
Outboard	Synchro	100	N/A	N/A
SDMS	8	1000	60	1000
SSDS	50$^{\text{Note 1}}$	20	10	10
SWAN	50$^{\text{Note 1}}$	20	10	10
TCS	8	20	10	10
TDBM	1	100	N/A	1000
TRD-F	1	100	N/A	N/A
WSC-3	1	N/A	N/A	N/A

Note 1 Multiple messages, highest rate given.

Note 2 Rate is manually selectable as 1, 8, 16, or 50 Hz.

DISPLAY AND CONTROL SUBSYSTEM (DCS)

The NAVSSI DCS Sensor Data Segment (SDS) and U.S. Coast Guard Integrated Navigation Segment (COMDAC-INS) CSCIs provide the integrated human–computer interface (HCI) for the NAVSSI program. The HCI is available from both the DCS workstation console and the NRS.

The DCS provides NAVSSI operator, ship's navigator, and navigation watch team with tools to plan, monitor, and carry out ownship's navigation; the capability to access and display digital nautical charts (DNCs); the capability to control and monitor RTS operations; a display of navigation sensor data; the capability to record and retrieve navigation information; and the capability to serve as the display and control unit for the GVRC installed in each RTS.

COMDAC-INS is integrated with the DCS SDS to provide a full set of automated navigation tools to the operator. The Joint Mapping Toolkit (JMTK) utilities provided with the DII COE provide the required chart manipulation functions. In combination, NAVSSI's CSCIs provide a navigation system in compliance with U.S. Navy policies and procedures for vessel navigation.

DCS Input from the RTS(s)

The DCS can input data from the RTS(s) via the NAVSSI LAN. The NAVSSI LAN is either the existing shipboard Fiber Data Distributed Interface (FDDI) LAN or a dedicated NAVSSI FDDI LAN. Input data includes external interface communication status, navigation sensor data, alarm, alert, and warning messages from the RTS(s). Input errors and out-of-tolerance conditions are displayed at the DCS and NRS.

Input from the Workstations

The DCS accepts manually input data from the DCS workstation and the NRS. The DCS provides displays to facilitate the manual input of system configuration data including the hardware suite designation and the sensor and user interfaces installed on that particular ship. This includes system configuration data, OSRP and lever arm data, and input of position fixes, course, and speed.

Digital Nautical Chart Access

The DCS HCI is capable of accessing and displaying any of the DNCs produced by the National Imagery Mapping Agency (NIMA) in a convenient and timely manner in accordance with U.S. Navy navigation policy.

Navigation Sensor Data Displays

The DCS provides the NAVSSI operator with the ability to view the most recent navigation data for each of the available sensors such as INSs, GPS receivers, speed logs, depth sensors, wind sensors, manually entered position, course, and speed.

Depending on the particular type of navigation sensor, the data to be displayed includes some subset of the following: time, latitude, longitude, velocity north, velocity east, total velocity (speed over ground), course over ground (COG), ship's heading, attitude data, estimated position (EP), fathometer depth data, magnetic bearing data, true and relative wind speed and direction, source selection algorithm status, interface status, true bearings, radar ranges, and lines of position.

NAVSSI Navigation Status Displays

The DCS provides continuously displayed navigation status lines. The navigation status lines are updated whenever the displayed values change up to a maximum rate of once per second. The data displayed on the status display lines include lines for the following: universal time coordinated (UTC), latitude, longitude, COG, speed over ground (SOG), soundings, data recording status, system alert status, system alarm status, navigation source selection mode (automatic or manual), and INS fix source selection mode.

User Interface Status Display

The DCS generates a display for simultaneous monitoring of the status of each sensor and user interface. The display lists each sensor and user interface and indicates whether that interface is active, i.e., ENABLED or DISABLED. For sensor interfaces, the display also shows whether or not that sensor is a current source for navigation data and whether its selection as a data source is automatic or manual. For user interfaces, the display shows the data source it is receiving and whether that data source was selected automatically or manually.

Navigation History Displays

The DCS can display navigation history data logs for the preceding 24 hours at UTC rollover each day. The data logs are as follows: position and depth provided by NAVSSI, position and depth provided by external users, ship's distance, source selections, GPS fix data from antenna No. 1 and from antenna No. 2 (if installed), fix data applied to INSs (if installed), manually entered fix data, courses and speeds, NAVSSI system crash data, alerts and alarms, and accumulated ownship distance.

DCS Controls

The DCS operator can control all the functions carried out by NAVSSI using graphical user interface (GUI)-based HCI control functions.

Navigation Source Selection Control

The DCS enables the NAVSSI operator to select an operating mode for the navigation source selection algorithms of "Automatic" or "Manual."

INS Fix Source Selection Control

The DCS enables the NAVSSI operator to choose what sensor NAVSSI will use to provide data to the ship's INSs (AN/WSN-5s or AN/WSN-7s).

RTS Expansion Port Controls

In addition to the standard control functions that the DCS performs for all RTS interfaces, the DCS provides the capability to rename expansion ports. Once renamed, the DCS uses the new name in all windows that include the expansion ports.

Time Source Selection Controls

The DCS enables the NAVSSI operator to select a source of time for the RTSs and it provides an actual time for the RTSs to correct to. There are two modes of time source selection: "Automatic" or "Manual." The default mode is the Automatic Mode.

The RTS does not select the time source if that time source does not meet the criteria checklist. The RTS continues to use whatever time source it had been using, whether it had been chosen manually or automatically. The RTS provides an alert to the DCS to advise the user that the chosen source was unacceptable. Faults and failures with the precise time distribution shall be reported to the DCS as alerts.

Navigation History Data Logging

The DCS records the following time tagged navigation history data collected during the preceding 24 hours to non-volatile storage: position and depth as indicated at the DCS (0.1-Hz rate); position and depth provided to external users (1-Hz rate); ship distance; source selection history; GPS fix data from antenna No. 1 and from antenna No. 2 (if installed) (1-Hz rate); fix data applied as input to the INSs; manually entered fix data; courses and speeds; and NAVSSI alerts and alarms.

All data except the 0.1-Hz DCS position and depth file are maintained for a minimum of 24 hours before being overwritten. The 0.1-Hz DCS position and depth data are maintained for at least 45 days before being overwritten.

Data to Tactical Database Manager

The DCS provides navigation data to the shipboard Tactical Database Manager (TDBM) at a user-selectable rate. The TDBM Application Programming Interface (API) calls are used by the DCS to provide these data to the track database via the FDDI LAN.

Supply Almanac Data to TAMPS

The DCS can provide GPS almanac data to the Tactical Air Mission Planning System (TAMPS). The DCS can transmit the almanac files to TAMPS via the LAN. In addition, the DCS can save the data to a 3.5-inch floppy diskette for direct use by TAMPS.

Digital Mapping, Charting, and Geodesy (MCG) Product Serving

NAVSSI can serve NIMA digital MCG products to DII COE shipboard user systems via local network connections. NAVSSI can also provide these digital products to non-DII COE clients by providing raw data via Network File System (NFS) and will provide User Datagram Protocol (UDP) broadcasts to notify users of updates that have been made by the operator.

GVRC Controls and Displays

The DCS serves as the control and display unit for the GVRCs.

COMDAC–INS

The U.S. Navy and Coast Guard implemented a joint development effort for chart display and manipulation called Command Display and Control–Integrated Navigation Segment (COMDAC-INS). COMDAC–INS is a DII segment that serves as the charting segment in the NAVSSI Block 3 DCS as defined in the NAVSSI-B3-IRS-101. The DCS HCI is developed under the DII COE architecture to work with the COMDAC–INS in providing the capabilities described in the following sections.

ECDIS-N and DoD Interoperability Requirements

The COMDAC–INS is designed in accordance with the guidelines and performance standards outlined in the ECDIS-N Policy Letter [2]. It accepts inputs from Navy standard automated and continuous positioning systems. The COMDAC–INS will accept radar and visual navigation fix information.

Display Requirements

The COMDAC–INS is designed to display all system digital nautical chart (SDNC) information, which is subdivided into three categories: standard display, display base, and all other information. When a chart is first displayed, it provides the standard display at the largest scale available.

The system will display DNC information and updates without degradation of their information content once the chart update format is determined by NIMA. A "north-up" orientation is required, with others permitted. The system uses recommended International Hydrographic Organization

(IHO) colors provided by NIMA symbology set (GEOSYM) and is visible in both day and night conditions. The system can display SDNC information for route planning, monitoring, and supplementary navigation tasks.

Display of Other Navigational Information

Radar and other navigational information may be added to the COMDAC-INS display. The system is designed not to degrade the SDNC information and to remain clearly distinguishable from the SDNC information.

Route Planning

The system can carry out route planning in a simple and reliable manner. It provides for route planning (including straight and curved segments) and route adjustment (e.g., adding, deleting, or changing the position of waypoints to a route). It is possible to plan an alternative route in addition to the selected route. The selected route is clearly distinguishable from the other routes. The mariner can specify a limit of deviation from the planned route. An automatic off-track alarm when deviating from a planned route by the limit specified is provided. An indication is provided if the mariner plans a route across an own ship's safety contour or the boundary of a prohibited area or of a geographical area for which special conditions exist.

Route Monitoring

The system can carry out route monitoring in a simple and reliable manner. For route monitoring, the selected route and ownship's position normally appear whenever the display covers that area. It is also possible to display a sea area that does not have the ship on the display (e.g., for look ahead, route planning). If this is done on the display in use for route monitoring, the automatic route monitoring functions are continuous (e.g., updating ship's position, providing alarms and indications). It is possible to return immediately to the route monitoring display covering ownship's position by single operator action.

The system provides an alarm within a specified time set by the mariner if the ship is going to cross the safety contour. It provides an alarm or indication if the ship is going to cross the boundary of a prohibited area or of a geographical area for which special conditions exist. An alarm is given when the specified limit for deviation from the planned route is exceeded. The system provides an indication when the input from the position-fixing system is lost, and it also repeats, but only as an indication, any alarm or indication passed to it from a position-fixing system. An alarm is given if the ship is going to reach a critical point on the planned route within a specified time or distance set by the mariner.

The system continuously plots the ship's position. It provides for the display of an alternative route in addition to the selected route. The selected route is clearly distinguishable from the other routes.

The system displays time-labels along the ship's track and other symbols required for navigation purposes such as the following: own-ship past track with time marks for both primary and secondary track; vector for course and speed made good; variable range marker and/or electronic bearing line; cursor; event posting for both DR position and time and EP and time; fix and time; position line time; transferred position line and time for both the predicted and actual tidal stream or current vector with effective time and strength (in box); danger highlight; clearing line; planned course and speed to make good; waypoint; distance to run; planned position with date and time; visual limits of lights arc to show rising/dipping; and position and time of "wheelover."

Voyage Recording

The system can reproduce certain minimum elements required to reconstruct the navigation history and verify the official database used during the previous 12 hours. The system records the complete track for the entire voyage at intervals not exceeding 4 hours. These data are protected and it is not possible to manipulate or change the recorded information.

STELLA

Another program integrated into the NAVSSI DCS is the System to Estimate Latitude and Longitude Astronomically (STELLA). STELLA is a software module developed to provide an integrated set of planning and reduction tools for celestial navigation. STELLA consists of the NAVSSI GUI and a U.S. Naval Observatory (USNO)-developed computational engine (CE). The GUI accepts input data and user commands, calls the CE function to process data, and displays the returned data in text, table, or graphics forms. The CE performs all necessary astronomical and navigational functions.

ACKNOWLEDGMENTS

The authors would to thank the many contributors to the NAVSSI system specification and project.

REFERENCES

1. System/Subsystem Specification for the Navigation Sensor System Interface (NAVSSI) AN/SSN-6 Block 3, NAVSSI-B3-SSS-SYS-REVBC0, December 1999.
2. Electronic Chart Display and Information System–Navy (ECDIS–N) Policy Ltr 3140 Ser N00/8U5000076, 17 March 1998.

❖

Peter Shaw

BS in Electrical Engineering, Temple University, 1983

Current Work: Deputy for Navigation and Digital Charting Systems in the Marine Navigation Division.

Bill Pettus

BS in Electrical Engineering, Rensselaer Polytechnic Institute, 1986

Current Work: Leads an Office of Naval Research program to improve the integration of relative navigation data available from tactical data links with the navigation suites on Navy platforms.

HMS *Scott*: United Kingdom Ocean Survey Ship

Fred Pappalardi, Steven J. Dunham, and Martin E. Leblang
SSC San Diego

ABSTRACT

Minimizing the cost per survey mile while ensuring that survey products meet required standards is a prime consideration when evaluating oceanographic surveying systems. This was one of the prime factors that led to the United Kingdom Ministry of Defense Procurement Executive (UK MOD PE) selection of a U.S. Navy designed ocean survey system to be installed aboard a new construction ship. The 13,500-ton HMS Scott *was designed and built specifically to accommodate the U.S. survey system and is considered the UK's premier survey ship. The mission of HMS* Scott *is to gather, process, and record time-correlated bathymetric, gravity, magnetic, and other oceanographic data as a function of latitude and longitude. Since deployment in early 1998, HMS* Scott *has successfully conducted highly accurate bathymetric surveys at an average sustained speed of 12 knots in ocean depths ranging from 50 fathoms to approximately 2500 fathoms in various types of terrain, from flat to very high relief.*

BACKGROUND

In 1963, the U.S. and the Government of the United Kingdom of Great Britain and Northern Ireland (UK) signed the Polaris Sales Agreement. The U.S. agreed to sell to the UK Polaris missiles (less warheads), equipment, and supporting services related to support of the UK Polaris submarine fleet.

As part of this support effort, the U.S. also shared, with the UK, ocean-bottom maps generated by a U.S. Navy developed wide-swath, multibeam bathymetric navigation system. This system, the first of its kind, was installed aboard three deep-ocean U.S. Navy survey ships and, for a period of more than 35 years, produced the highly accurate bathymetric charts required by the U.S. and UK Fleet Ballistic Missile (FBM) submarines.

In 1987, the UK Ministry of Defense (MOD) decided to update its ocean-surveying capability. After evaluating several candidate systems, the MOD concluded that the only system capable of meeting its FBM submarine requirements was the system developed and used by the U.S. Navy. Therefore, in 1995, the UK MOD approved construction of the 13,500-ton HMS *Scott*, shown in Figure 1, and specified that the ship was to be designed and built specifically to accommodate the U.S. Navy developed, fully integrated Ocean Survey System (OSS).

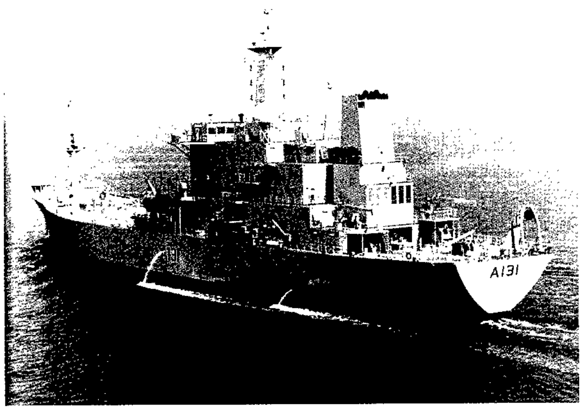


FIGURE 1. HMS *Scott*.

HMS *SCOTT* DESCRIPTION

HMS *Scott*, built by Appledore Shipbuilders Ltd. of North Devon, England, was handed over to the Royal Navy in 1997. She was designed to operate in areas remote from normal shipping lanes, in changeable severe weather conditions, and on a schedule of nine 35-day cycles each year, surveying for 24 hours a day when tasked [1]. Table 1 shows HMS *Scott* principal characteristics.

TABLE 1. HMS *Scott* principal characteristics.

Length	130 meters
Beam	21.5 meters
Design survey draft	8.3 meters
Displacement	13,300 tons
Survey speed	15 knots
Machinery plant	Diesel/Single screw
Endurance	35 days at survey speed
Thrusters	Bow
Crew	65

Figure 2 is a profile drawing of HMS *Scott* showing locations of mission-related spaces. Within these spaces are housed the elements that make up the OSS that will be discussed in this paper.

MISSION SYSTEM DESCRIPTION

The mission of HMS *Scott* is to gather, process, and record time-correlated bathymetric, gravity, and other ocean-related data. This paper will address only the mission of gathering bathymetric data, as accomplished by the OSS.

The entire OSS is supported by a dedicated, regulated power system, also developed by the U.S. Navy. Figure 3 is a simplified block diagram of the OSS.

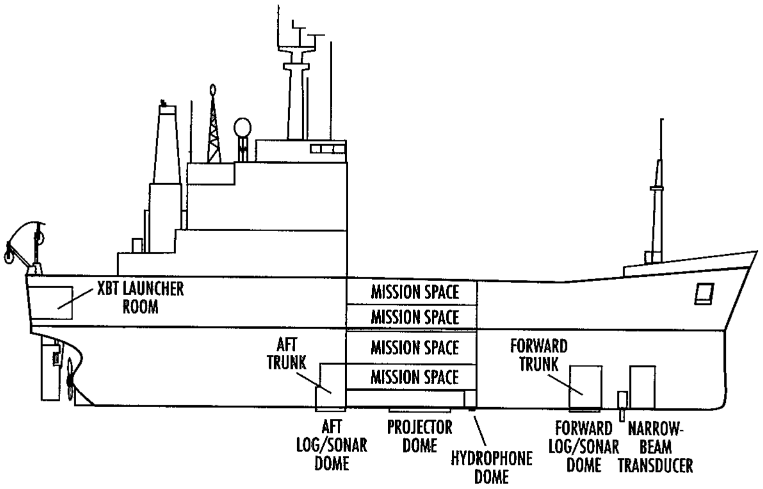


FIGURE 2. HMS *Scott* profile.

NAVIGATION SUBSYSTEM

The Navigation Subsystem provides precise and accurate platform attitude, position, and velocity information as a function of time for correlation with depth and other recorded survey data to produce specialized charts. Figure 4 is a simplified functional block diagram of the Navigation Subsystem.

The heart of the Navigation Subsystem is the navigation computer that performs the data integration, monitoring, and control functions that coordinate overall operation of the Navigation Subsystem.

Position data from Loran-C receivers, the Global Positioning System (GPS), and the Miniature Ship's Inertial Navigation System (MINISINS) are prioritized and filtered by the navigation computer program to produce the best present position (BPP) output. BPP is supplied in terms of latitude and longitude to the Sound Velocity System (SVS), Gravity System, and the Mission Control and Processing Subsystem (MCAPS). The MCAPS consists of the Survey Control System (SCS), System Analysis Station (SAS), and the Data Refinement System (DRS).

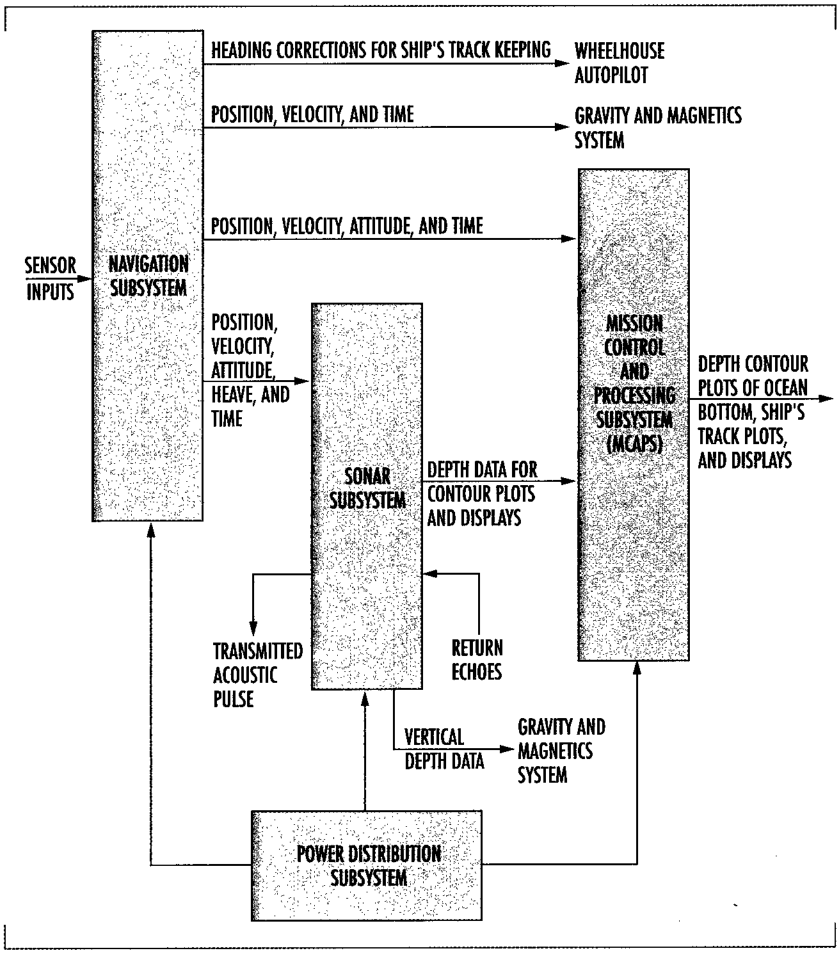

HEADING CORRECTIONS FOR SHIP'S TRACK KEEPING
WHEELHOUSE AUTOPILOT
POSITION, VELOCITY, AND TIME
GRAVITY AND MAGNETICS SYSTEM
POSITION, VELOCITY, ATTITUDE, AND TIME
SENSOR INPUTS
NAVIGATION SUBSYSTEM
POSITION, VELOCITY, ATTITUDE, HEAVE, AND TIME
MISSION CONTROL AND PROCESSING SUBSYSTEM (MCAPS)
DEPTH CONTOUR PLOTS OF OCEAN BOTTOM, SHIP'S TRACK PLOTS, AND DISPLAYS
SONAR SUBSYSTEM
DEPTH DATA FOR CONTOUR PLOTS AND DISPLAYS
TRANSMITTED ACOUSTIC PULSE
RETURN ECHOES
VERTICAL DEPTH DATA
GRAVITY AND MAGNETICS SYSTEM
POWER DISTRIBUTION SUBSYSTEM

FIGURE 3. Simplified Ocean Survey System block diagram.

Velocity data, including vertical velocity, are received by the navigation computer program, and those inputs from the selected velocity source are distributed to the wide-swath-array sonar and to the Gravity System.

Attitude data are provided from both the MINISINS and the MK-29 Gyrocompass, a backup to the MINISINS. Source selection and distribution of the data are accomplished by the Ship's Attitude Data Converter (SADC). Heave data, developed in the heave processor, are distributed along with attitude data.

Roll, pitch, heading, and heave data are supplied to the Sonar Subsystem. Roll, pitch, and heading are also provided to the navigation computer to compensate for antenna lever arms to the ship's reference point. Heading is supplied to the shipboard heading indicators, ship's auto-pilot, and the SCS.

Heading corrections are developed by the navigation computer program and supplied to the ship's auto-pilot for use during track-keeping operations. The heading corrections are combined with the auto-pilot steering commands and cause the ship to steer over a prescribed ground track instead of steering to a prescribed heading. This capability provides high-quality, tight track control.

SONAR SUBSYSTEM

The Sonar Subsystem consists of a wide-swath-array sonar and a single-beam sonar. Figure 5 is a simplified block diagram of the entire Sonar Subsystem. The wide-swath-array sonar obtains rapid acquisition of high-resolution sonar data as a function of time as the ship progresses along a desired track. The system employs a 120-degree, fan-shaped acoustic swath pattern that is transverse to the ship's track (see Figure 6) and operates in depths ranging from 50 fathoms to beyond 6000 fathoms.

To obtain depth measurements, the system transmits a 7-ms pulse every 12 or 15 seconds in deep water, and a 3-ms pulse every 3 or 6 seconds in shallow water. The transition from 12 to 15 seconds takes place at approximately 2000 to 2400 fathoms to allow for increased processing time. The transmit frequency is 12 kHz with each ping consisting of a 7-ms pulse. The returning echoes, received by the 144 hull-mounted hydrophones, are sampled every 3-ms to give time snapshots of the

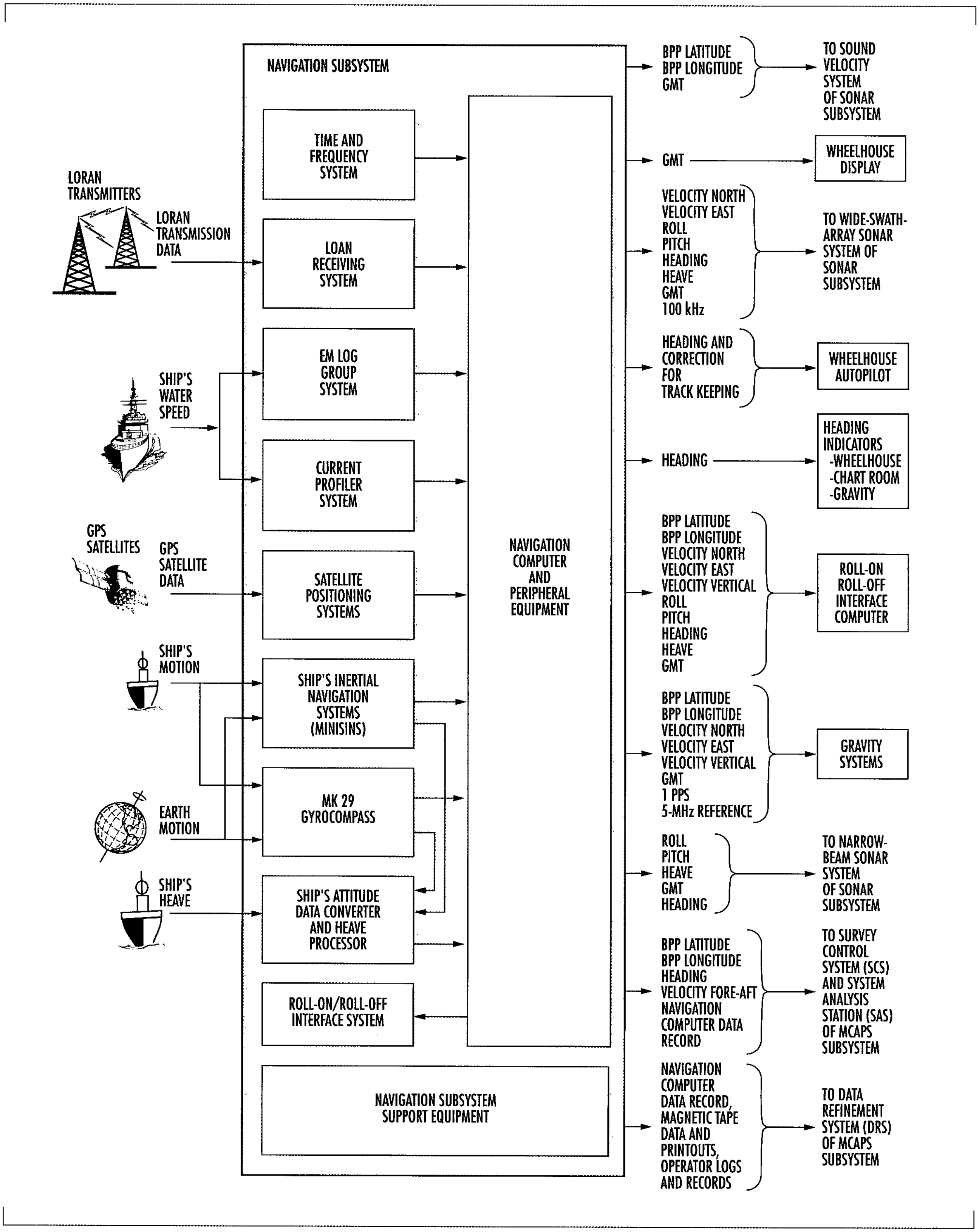

NAVIGATION SUBSYSTEM
TIME AND FREQUENCY SYSTEM
LORAN TRANSMITTERS
LORAN TRANSMISSION DATA
LOAN RECEIVING SYSTEM
EM LOG GROUP SYSTEM
SHIP'S WATER SPEED
CURRENT PROFILER SYSTEM
GPS SATELLITES
GPS SATELLITE DATA
SATELLITE POSITIONING SYSTEMS
SHIP'S MOTION
SHIP'S INERTIAL NAVIGATION SYSTEMS (MINISINS)
EARTH MOTION
MK 29 GYROCOMPASS
SHIP'S HEAVE
SHIP'S ATTITUDE DATA CONVERTER AND HEAVE PROCESSOR
ROLL-ON/ROLL-OFF INTERFACE SYSTEM
NAVIGATION COMPUTER AND PERIPHERAL EQUIPMENT
NAVIGATION SUBSYSTEM SUPPORT EQUIPMENT
BPP LATITUDE
BPP LONGITUDE
GMT
TO SOUND VELOCITY SYSTEM OF SONAR SUBSYSTEM
GMT
WHEELHOUSE DISPLAY
VELOCITY NORTH
VELOCITY EAST
ROLL
PITCH
HEADING
HEAVE
GMT
100 kHz
TO WIDE-SWATH-ARRAY SONAR SYSTEM OF SONAR SUBSYSTEM
HEADING AND CORRECTION FOR TRACK KEEPING
WHEELHOUSE AUTOPILOT
HEADING
HEADING INDICATORS
-WHEELHOUSE
-CHART ROOM
-GRAVITY
BPP LATITUDE
BPP LONGITUDE
VELOCITY NORTH
VELOCITY EAST
VELOCITY VERTICAL
ROLL
PITCH
HEADING
HEAVE
GMT
ROLL-ON ROLL-OFF INTERFACE COMPUTER
BPP LATITUDE
BPP LONGITUDE
VELOCITY NORTH
VELOCITY EAST
VELOCITY VERTICAL
GMT
1 PPS
5-MHz REFERENCE
GRAVITY SYSTEMS
ROLL
PITCH
HEAVE
GMT
HEADING
TO NARROW-BEAM SONAR SYSTEM OF SONAR SUBSYSTEM
BPP LATITUDE
BPP LONGITUDE
HEADING
VELOCITY FORE-AFT
NAVIGATION COMPUTER DATA RECORD
TO SURVEY CONTROL SYSTEM (SCS) AND SYSTEM ANALYSIS STATION (SAS) OF MCAPS SUBSYSTEM
NAVIGATION COMPUTER DATA RECORD, MAGNETIC TAPE DATA AND PRINTOUTS, OPERATOR LOGS AND RECORDS
TO DATA REFINEMENT SYSTEM (DRS) OF MCAPS SUBSYSTEM

FIGURE 4. Navigation Subsystem.

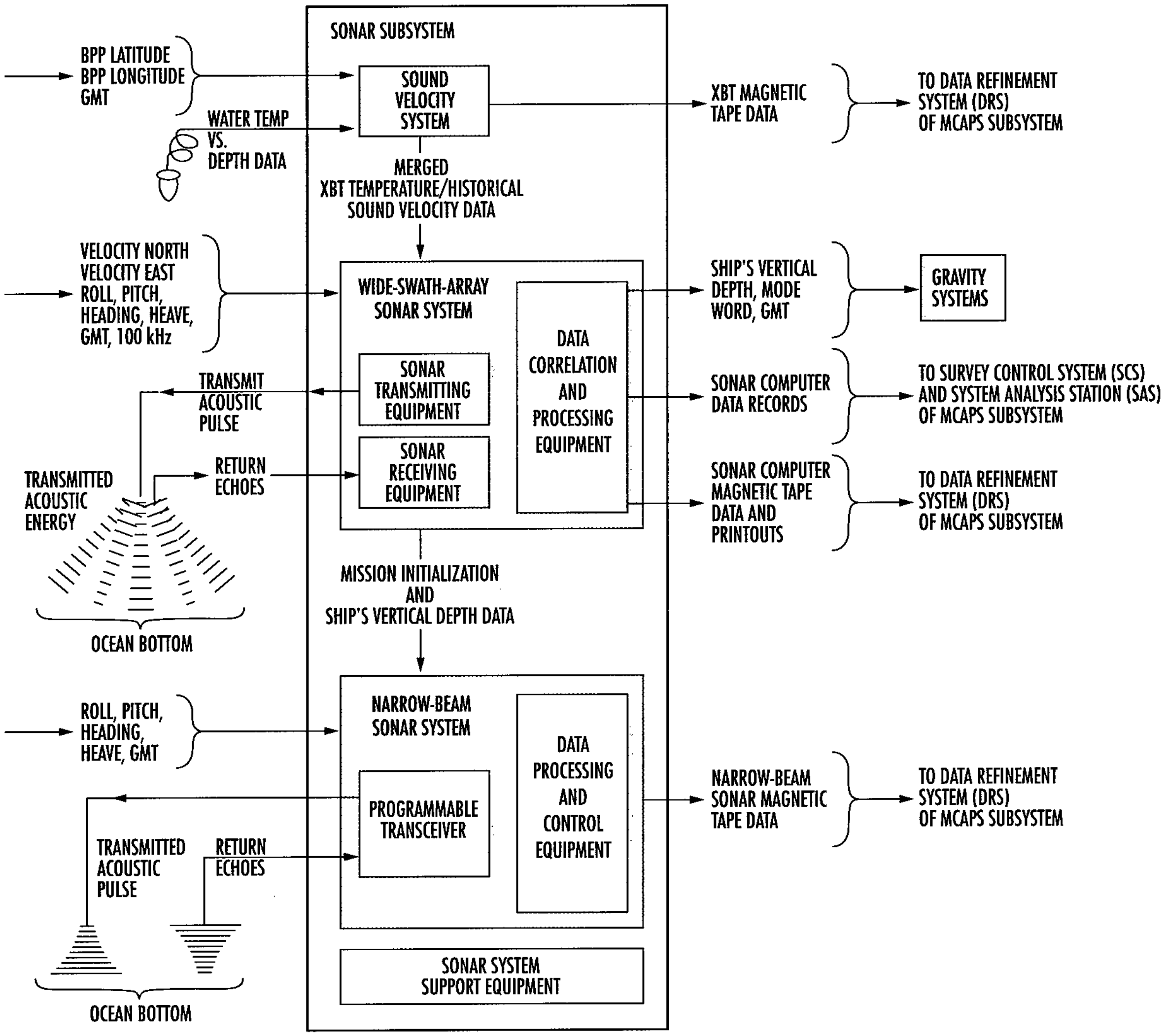


FIGURE 5. Sonar Subsystem.

acoustic pulse reflected from the bottom. Digital signal-processing algorithms process up to 1664 time snapshots of the returning signal to develop an across-track profile of the bottom, with an average internal resolution of a third of a degree across the swath. The profile is then decimated to 121 bottom points spaced 1 degree apart. Each point is then corrected for sound-velocity variations and for roll, pitch, heading, and heave variations between time of transmission and reception. The data are then checked for validity and reasonableness by the on-line software.

Transmission takes place with the projection of a burst of acoustic energy (from 58 projectors mounted in the fore-aft direction along the ship's centerline) that ensonifies a narrow strip of the ocean floor. The width of the ensonified strip extends beyond the 120-degree swath pattern. Compensation for pitch angles of up to ±10 degrees is applied during transmission to steer the projected acoustic energy beam to the vertical about the pitch axis. Acoustic echoes returning within the 120-degree swath pattern are captured by an array of 144 hydrophones (mounted athwartship and forward of the projector array) for processing. The swath pattern is roll-corrected to the vertical. Compensation for roll angles of up to ±15 degrees is applied to returns upon reception to keep the swath pattern fixed relative to the vertical about the roll axis.

Figure 7 shows wide-swath-array sonar depth versus bottom coverage. For depths down to 1000 fathoms, each transmission, or ping, provides 121 depth points over a data-acquisition swath that is 3.5 miles wide. As depths go beyond 1000 fathoms, the number of measurable data points captured within the 120-degree swath pattern gradually begins to diminish. At a depth of 5000 fathoms, each ping provides 91 depth points over a data-acquisition swath that is 12 miles wide. To allow for the varied acoustic-signal travel times associated with shallow and deep depths, the system employs operator selected ping rates of 3, 6, 12, or 15 seconds.

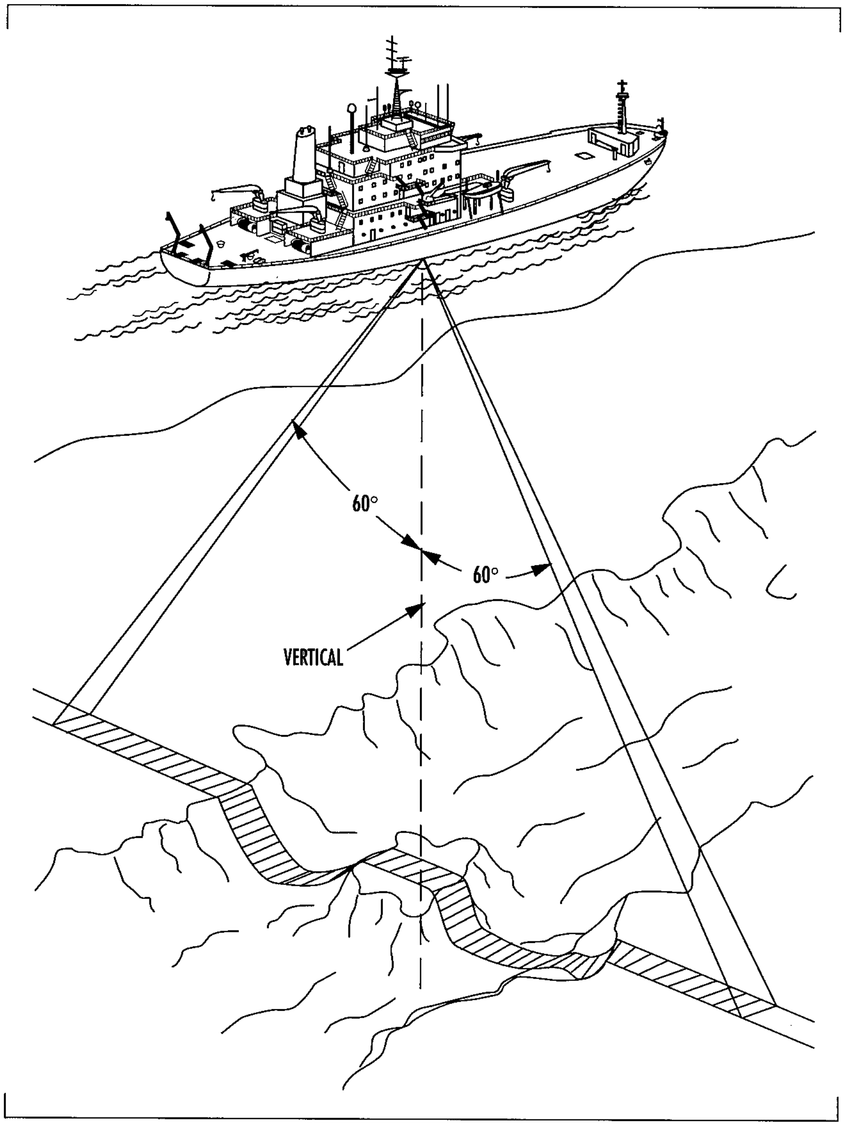


FIGURE 6. Wide-swath-array sonar beam pattern.

Sound-velocity corrections are applied by using periodic expendable bathythermograph (XBT) casts to measure the ocean water temperature as a function of depth. These water temperature versus depth measurements provide the means to detect a significant change in the sound-velocity structure of the local ocean area of interest and to determine the applicability of the sound-velocity-versus-depth profile in current use by the Sonar Subsystem. The XBT cast data are merged with historical sound-velocity data and sent to the Sonar Subsystem for use during returned echo data processing.

The single-beam sonar uses a 9-degree conical pattern (see Figure 8) and obtains and records the ocean depth directly beneath the ship. This sonar provides an independent method of monitoring the vertical-depth data output of the wide-swath-array sonar. When the wide-swath sonar is off-line, such as when the ship is in water less than 50 fathoms or when it is inoperative, the single-beam sonar performs the depth-acquisition function of the OSS.

MISSION CONTROL AND PROCESSING SUBSYSTEM

The MCAP Subsystem (Figure 9) provides centralized control and performance monitoring of overall OSS operation. The MCAP Subsystem consists of three systems: Survey Control System (SCS), Systems Analysis Station (SAS), and Data Refinement System (DRS), all interconnected over a Local Area Network (LAN).

SURVEY CONTROL SYSTEM

The SCS provides remote-operator control functions for the navigation and sonar computers along with graphical data displays necessary to support on-line survey operations. The SCS provides a centralized, on-line, survey-system control workstation that provides the following four major functions: (1) initialization of the navigation and wide-swath-array sonar computer programs, (2) performance of system mode changes, resets, and parameter updates, (3) survey data collection, and (4) displays of graphical and tabular data providing high-level representations of the current navigation and sonar data being collected, as well as a display of ship's progress along the prescribed survey track.

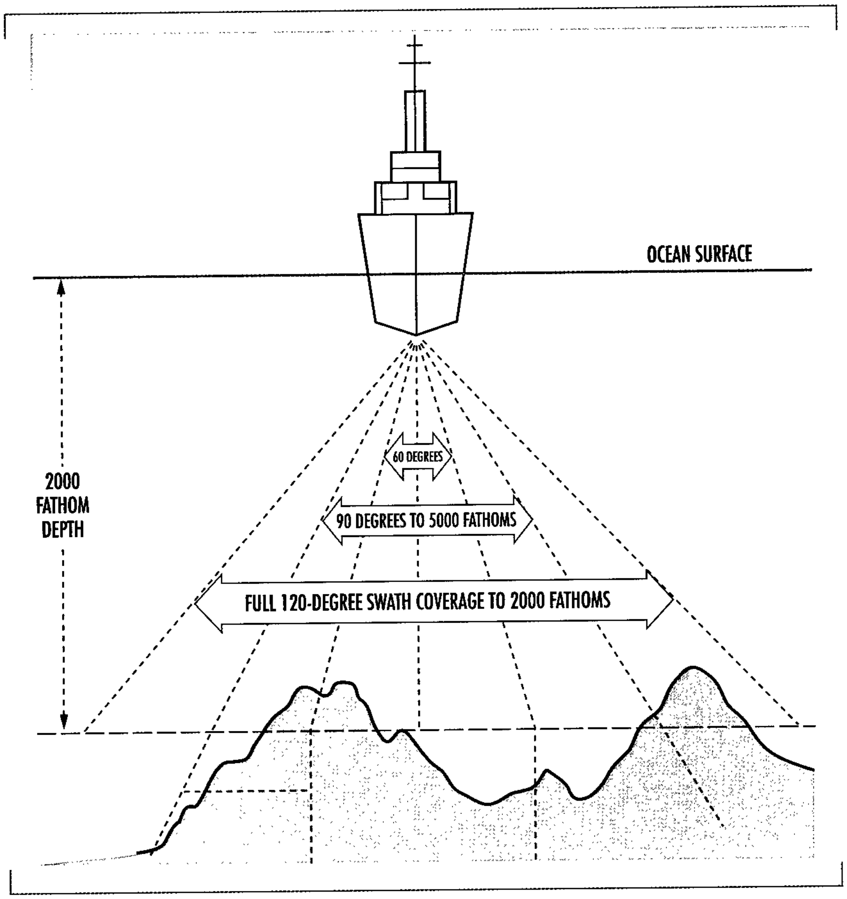

OCEAN SURFACE
2000 FATHOM DEPTH
60 DEGREES
90 DEGREES TO 5000 FATHOMS
FULL 120-DEGREE SWATH COVERAGE TO 2000 FATHOMS

FIGURE 7. Wide-swath-array sonar coverage vs. depth.

SCS display functions include the high-level graphical and tabular data displays necessary for the operator to quickly assess and ensure that data collected by the Navigation and Sonar Subsystems meet survey mission requirements.

SYSTEM ANALYSIS STATION

The SAS serves as the central workstation for control of plotting functions, and on-line system performance analysis, monitoring, and troubleshooting through a comprehensive set of graphic displays. Some of the graphic displays that the operator can select to examine system performance, both in near-real-time or post-time are plots of position, velocity, ship's track, or depth contours. A database of up to 50 days' worth of data can be accessed by the operator for review and analysis.

DATA REFINEMENT SYSTEM

The DRS provides post-time data refinement functions for data collected by the OSS. This off-line, computer-based processing system post-time processes shipboard-gathered sonar and navigation data and produces the final, fleet-ready bathymetric chart product. Specifically, the DRS develops refined position and velocity navigational data; generates hard-copy and screen displays of bathymetric navigation charts; and edits, analyzes, displays, and compresses the wide-swath sonar data.

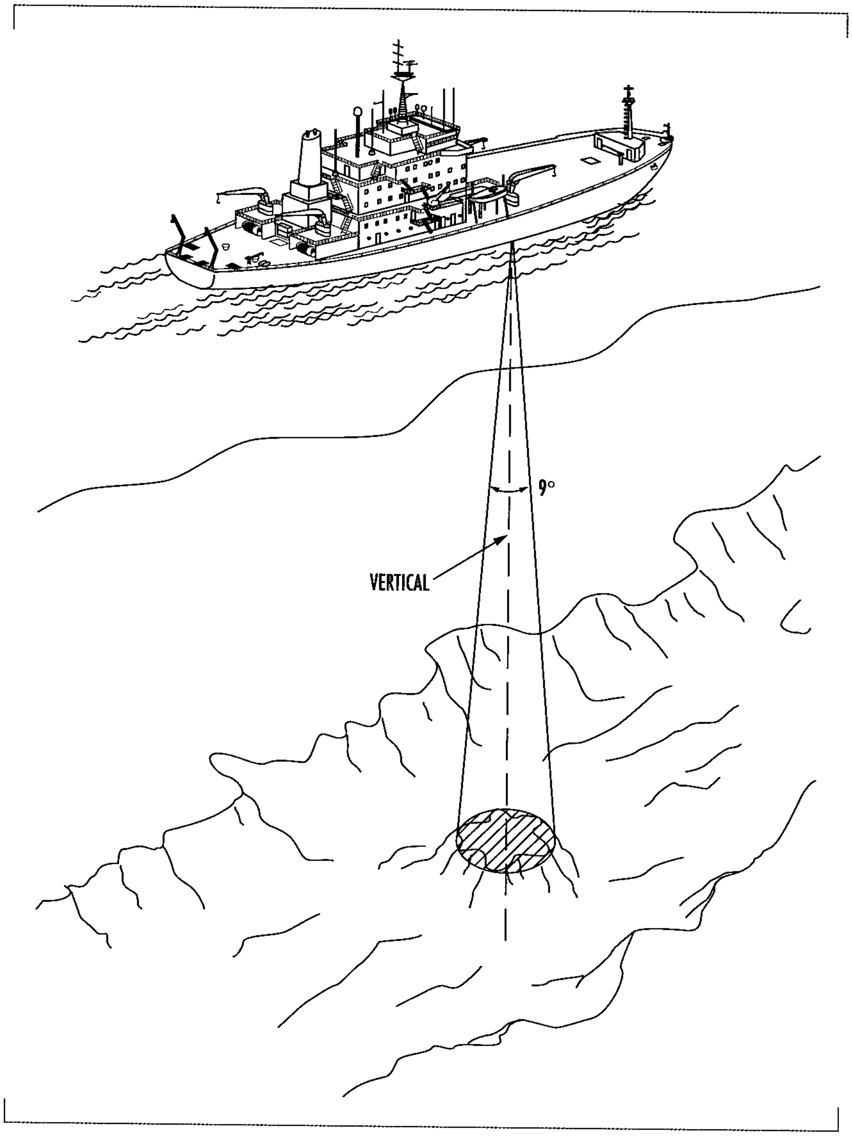


FIGURE 8. Single-beam sonar system beam pattern.

POWER DISTRIBUTION SYSTEM

The primary function of the Power Distribution System (Figure 10) is to provide precision-regulated and signal-conditioned 60- and 400-Hz power to all OSS equipment. A secondary, but equally important function is to maintain a continuous emergency back-up power supply for critical equipment if the main power supply should fail. The 60- and 400-Hz power distribution systems receive power from the ship's service diesel generators via the ship's main service power panel.

The 60-Hz power distribution system provides 120-V, 60-Hz, three-phase power to the 60-Hz regulated power panels of the OSS. The 60-Hz power distribution system consists of two regulated power systems and one uninterruptible power source (UPS). The 400-Hz power distribution system provides 120-V, 400-Hz, three-phase delta output power to the OSS. The 400-Hz power distribution system consists of two separate and independent systems. Each system uses one 400-Hz UPS equipment cabinet and one 400-Hz UPS battery cabinet. Only one 400-Hz system is on at a time. The second system is maintained in an active standby operating mode.

The alarm and monitoring equipment associated with the Power Distribution System consists of the following items: (1) a power disturbance analyzer that monitors, measures, records, analyzes, and prints the type of disturbances that occur in the 60- and 400-Hz power distribution systems; (2) 60- and 400-Hz alarm and monitor panels that monitor AC voltages, AC current, and frequencies of the vital equipment power panels; (3) 60- and 400-Hz UPS remote status panels that monitor and provide a visual operating status of the UPS systems including battery condition and that provide visual and audible alarms for system changes.

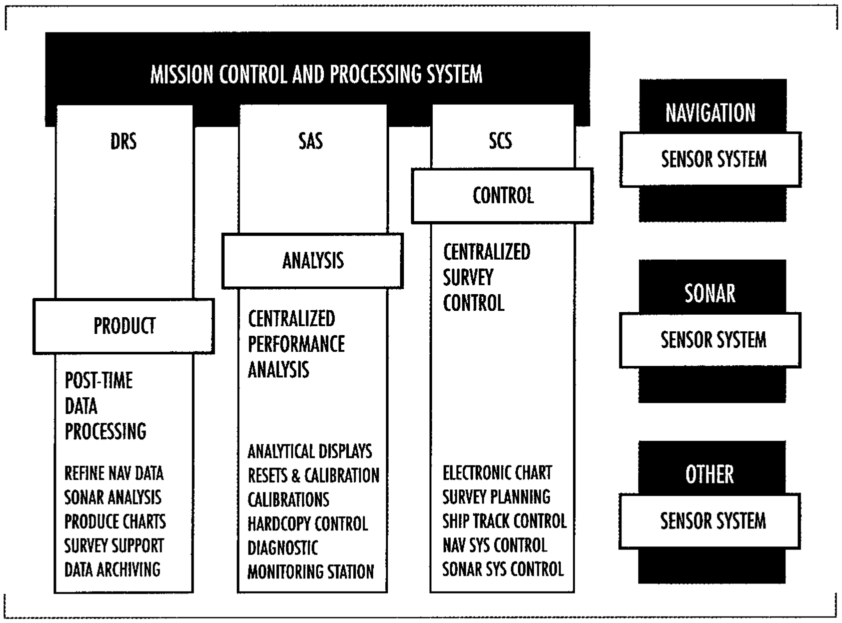


FIGURE 9. Mission control and processing system.

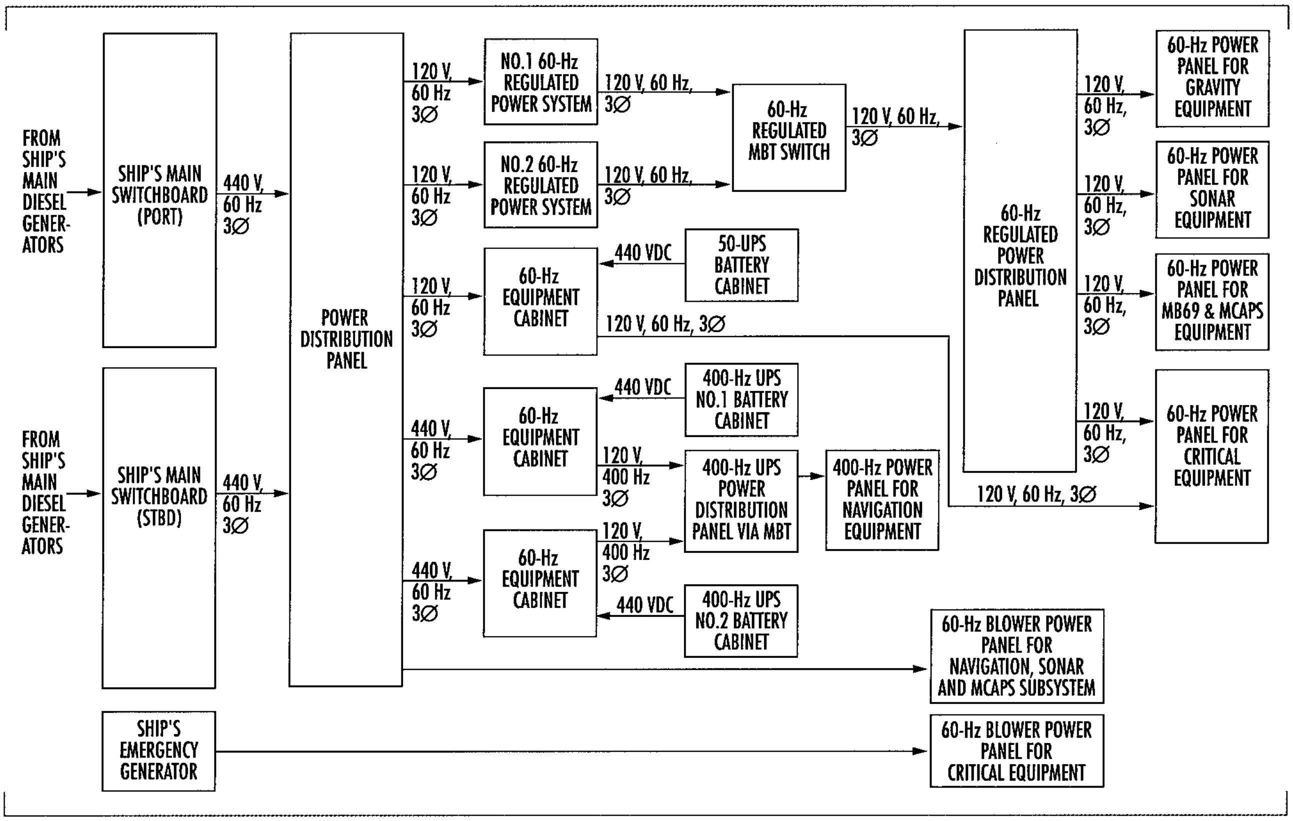

FROM SHIP'S MAIN DIESEL GENER-ATORS
SHIP'S MAIN SWITCHBOARD (PORT)
440 V, 60 Hz 3∅
FROM SHIP'S MAIN DIESEL GENER-ATORS
SHIP'S MAIN SWITCHBOARD (STBD)
440 V, 60 Hz 3∅
SHIP'S EMERGENCY GENERATOR
POWER DISTRIBUTION PANEL
120 V, 60 Hz 3∅
NO.1 60-Hz REGULATED POWER SYSTEM
120 V, 60 Hz, 3∅
120 V, 60 Hz 3∅
NO.2 60-Hz REGULATED POWER SYSTEM
120 V, 60 Hz, 3∅
60-Hz REGULATED MBT SWITCH
120 V, 60 Hz, 3∅
120 V, 60 Hz 3∅
60-Hz EQUIPMENT CABINET
440 VDC
50-UPS BATTERY CABINET
120 V, 60 Hz, 3∅
440 V, 60 Hz 3∅
60-Hz EQUIPMENT CABINET
440 VDC
400-Hz UPS NO.1 BATTERY CABINET
120 V, 400 Hz 3∅
400-Hz UPS POWER DISTRIBUTION PANEL VIA MBT
400-Hz POWER PANEL FOR NAVIGATION EQUIPMENT
440 V, 60 Hz 3∅
60-Hz EQUIPMENT CABINET
120 V, 400 Hz 3∅
440 VDC
400-Hz UPS NO.2 BATTERY CABINET
60-Hz REGULATED POWER DISTRIBUTION PANEL
120 V, 60 Hz, 3∅
60-Hz POWER PANEL FOR GRAVITY EQUIPMENT
120 V, 60 Hz, 3∅
60-Hz POWER PANEL FOR SONAR EQUIPMENT
120 V, 60 Hz, 3∅
60-Hz POWER PANEL FOR MB69 & MCAPS EQUIPMENT
120 V, 60 Hz, 3∅
60-Hz POWER PANEL FOR CRITICAL EQUIPMENT
120 V, 60 Hz, 3∅
60-Hz BLOWER POWER PANEL FOR NAVIGATION, SONAR AND MCAPS SUBSYSTEM
60-Hz BLOWER POWER PANEL FOR CRITICAL EQUIPMENT

FIGURE 10. Power Distribution System.

CONCLUSION

The HMS *Scott* OSS described in this paper has been in operation since January 1998 and has provided the UK with a large volume of exceptionally accurate and detailed ocean-bottom data. At an average speed of 12 knots, approximately 10,000 survey miles can be achieved during a 35-day survey operation. Since the proven reliability of the OSS is greater than 99 percent, and given a specified minimum ship life of 25 years, it is expected that HMS *Scott* will provide large volume, very cost-effective, highly reliable, and very accurate oceanographic and bathymetric data over its operational lifetime.

ACKNOWLEDGMENTS

The authors would like to thank Mr. Gordon Smith, United Kingdom Ministry of Defense, who provided key information critical to the preparation of this paper and Eric Matsuo, SSC San Diego, for his invaluable assistance in formatting this paper.

AUTHORS

Steven J. Dunham

BS in Electrical Engineering, State University of New York at Buffalo, 1971

Current Research: Deep ocean survey system, HMS *Scott*; data refinement system, UK Hydrographic Office.

Martin E. Lablang

MS in Electrical Engineering, New York University, 1971

Current Research: Navigation system, HMS *Scott*.

Fred Pappalardi

BA in Electrical Engineering, Pratt Institute, New York City, 1963

Current Work: Business Manager for the Marine Navigation Division; development of new business opportunities in marine navigation.

REFERENCE

1. *Shipping World and Shipbuilding*, 1997 (May).

❖

The Use of Field Screening or Rapid Sediment Characterization (RSC) Tools for Sediment Assessments

Victoria J. Kirtay and Sabine E. Apitz
SSC San Diego

ABSTRACT

This paper discusses several rapid site characterization (RSC) technologies that can be used at marine sediment sites, including X-ray fluorescence (XRF) for metals, ultraviolet fluorescence (UVF) for polycyclic aromatic hydrocarbons, QwikSed bioassay for assessing toxicity, and other techniques. Examples are provided to illustrate the efficacy of applying RSC tools to different stages of the Ecological Risk Assessment (ERA) process. Finally, recommendations are given for the evaluation, selection, and application of RSC tools for the ERA process.

INTRODUCTION

The primary goals of a sampling and analysis plan for an Ecological Risk Assessment (ERA) or a sediment site assessment are to identify potential contaminant sources and to delineate areas of concern. However, traditional sampling and analysis approaches do not always provide the information necessary to support the decision-making process in a cost- and time-effective manner. Site assessments performed in the marine environment are often hindered due to the complexity and heterogeneity of marine ecosystems. Because of the complex nature of marine ecosystems, U.S. Navy policy [1] specifically requires that sampling programs focus primarily on the identification of potential contaminant sources and on the delineation of areas of contaminated media. Navy policy further dictates that sampling programs should make use of advanced chemical and biological screening technologies, data quality objectives, and statistical procedures to minimize overall sampling requirements. Implementation of advanced chemical, physical, and/or biological screening technologies (i.e., rapid sediment characterization tools) at different stages of the ERA process can aid in focusing sampling requirements and can ultimately facilitate reaching final decisions.

WHAT IS RAPID SEDIMENT CHARACTERIZATION?

Rapid sediment characterization (RSC) tools are field-transportable analytical tools that provide measurements of chemical, physical, or biological parameters on a near real-time basis. A variety of tools exist that are capable of making these types of measurements. Many technologies have been used to characterize different types of environmental media (e.g., soil, sediment, water, and air). These technologies are described in several Environmental Protection Agency (EPA) documents [2 and 3], including the online Field Analytical Technologies Encyclopedia (FATE) [4]. This encyclopedia provides information about technologies that can be used in the field to characterize contaminated soil and ground water, to monitor the progress of remedial efforts, and in some cases, to confirm sampling and analysis for site closeout. Although not all of the technologies currently available are applicable to sediment sites, several have been tested and demonstrated at Navy marine sediment sites (Table 1). Examples can also be found in standard environmental textbooks such as Gilbert's 1987 *Statistical Methods for Environmental Pollution Monitoring*, which provides specific examples of the use of screening and laboratory data together to optimize for reduction in cost and data variability [5].

WHY IS RCS IMPORTANT?

TABLE 1. Examples of rapid sediment characterization tools tested in marine sediments.

Analytical Technique	Parameter(s)
X-ray Fluorescence (XRF) Spectrometry	Metals (e.g., Cu, Zn, Pb)
UV Fluorescence (UVF) Spectroscopy	Polycyclic Aromatic Hydrocarbons (PAHs)
Immunoassays	Polychlorinated biphenyls (PCBs), PAHs and Pesticides
QwikSed Bioassay Microtox	Acute and Chronic Toxicity Acute Toxicity
Laser Particle Scattering	Grain Size (% fines)

An ERA evaluates the likelihood that exposure to one or more stressors (e.g., contaminants) will result in adverse ecological effects [6]. The purpose of the assessment is to provide information relevant to the management decision-making process. Ideally, ERAs should be scientifically based, defensible, cost-effective, and protective of human health and the environment (see, for example, [1]). Collection of data necessary to support decisions at sediment sites in a cost-effective manner is often hindered by the complexity and heterogeneity of marine ecosystems. Detailed site investigations require extensive sampling and subsequent laboratory analyses for both metal and organic contaminants. Samples are often collected without any *a priori* knowledge of the nature and extent of contamination. Because of the high cost of laboratory analyses, the number of samples taken is often cost-limited. Thus, zones of contamination can be missed or, if located, overestimated or underestimated. To obtain more detailed spatial information on the extent of contamination, researchers must often sample and analyze sites of interest in an iterative manner. Chemical assays are often combined with additional laboratory analyses, including one or several bioassays to determine whether there are adverse biological effects of these contaminants in various media (e.g., sediment, elutriate, water column). This approach can be prohibitively costly, slow, and labor-intensive. When used appropriately, RSC tools can streamline many aspects of the ERA process, delineating areas of concern, filling information gaps, and ensuring that expensive, certified analyses have the highest possible impact.

To determine if RSC tools are appropriate to assess contamination at a given site, several questions should be asked. For example: What are the goals of the investigation? What are the contaminants of concern? Are the contaminants known? What are the action limits? What are the strengths and weaknesses of the analytical methods being considered? Do instrument detection limits meet action limit requirements? By asking these questions before sampling begins and by considering the advantages and disadvantages of different techniques, appropriate decisions can be made on how best to implement a technology or suite of technologies to facilitate the ERA process. Table 2 lists the relative advantages and limitations of RSC methods and standard methods. A brief description of some RSC technologies that have been tested in sediments is provided below. All of these technologies described are commercially available.

TABLE 2. Advantages and limitations of screening and standard laboratory methods.

RCS Analysis	Standard Laboratory Analysis
Benefits · rapid results can guide sampling locations · potential for high data density for mapping · reduced cost per sample	Benefits · standard methods that are very quantitative · can often remove interferences
Limitations · often non-specific · semi-quantitative · matrix sensitive	Limitations · often blind-sampling · long delays to results · expensive ($K/sample)

EXAMPLES OF RSC TECHNOLOGIES: GENERAL PRINCIPLES

X-ray Fluorescence Spectrometry Metals

Commercially available, portable X-ray fluorescence (XRF) spectrometry analytical instruments can provide rapid, multi-element analysis of metals in sediment. Samples are exposed to X-ray energy, which liberates electrons in the inner shell of metal atoms. As the outer electrons cascade toward the inner shells to fill the vacancies, energy is released, or fluoresced. The fluorescing energy spectrum identifies the metals and each peak's intensity is proportional to concentration. Generally, XRF can detect and quantify elements from sulfur to uranium. For common metals, such as lead, zinc and copper, this method yields a detection limit range from 50 to 150 parts per million (ppm) and requires 2 to 5 minutes per analysis in soils and sediments. Commercial XRF instruments are readily available for purchase (~ $11,000 to $56,000) or lease (~ $150/day to $6000/month) depending on options and equipment size required. To accommodate field application, many instruments weigh less than 30 pounds and can be operated with batteries for 8 to 10 hours [4 and 7].

Ultraviolet Fluorescence Spectroscopy: PAHs

Fluorescence is a standard analytical technique that can be used to measure the concentration of various analytes in different matrices. Ultraviolet fluorescence spectrometry (UVF) can be used for the determination of polycyclic aromatic hydrocarbons (PAHs) in sediments. This technique is based on the measurement of fluorescence observed following UV excitation of either bulk samples or organic solvent extracts of sediments. However, detection limits are greatly enhanced by extraction. When UV light is passed through a sample, the sample emits light (fluorescence) proportional to the concentration of the fluorescent molecules (e.g., PAHs) in the sample [8]. An analysis, with extraction, can be done in 10 to 30 minutes, and for PAHs, the range for detection limits when using UVF is from 1 to 5 ppm total solid phase. UVF instruments are commercially available from various vendors for purchase (~ $10,000 to $12,000) or for weekly rental.

Immunoassays: PCBs, PAHs, Pesticides

This technique can be used for the identification and quantification of many organic compounds (e.g., polychlorinated biphenyls [PCBs], PAHs, and pesticides). Immunoassays use antibodies that have been developed to bind with a target compound or class of compounds. Concentrations of analytes are identified through the use of a sensitive colorimetric reaction. The determination of the target analyte's presence is made by comparing the color developed by a sample of unknown concentration with the color developed by a standard containing the analyte at a known concentration. The concentration of the analyte is determined by the intensity of color in the sample and is measured through use of a spectrophotometer. Immunoassay kits are relatively quick and simple to use. Several test kits are commercially available and range in cost from $10 to $40 per sample test kit. Detection limits can vary, depending on the dilution series used. For example, the detection limit for PCBs in sediments ranges from 50 to 500 parts per billion (ppb) [4 and 9].

Screening Bioassay Tests

The Microtox bioassay is a commercial test that measures the inhibition of light emitted by a bioluminescent microorganism. Any decrease in light output relative to controls suggests bioavailable contaminants or other stressors. Several studies have compared Microtox response to other bioassays (e.g., [10]).

The QwikSed rapid bioassay system is proving to be a valuable asset for conducting bioassays on marine sediments. The basis of detection is to measure a reduction in light from a bioluminescent dinoflagellate such as *Gonyaulux polyedra* or *Ceratocorys horrida* following exposure to a toxicant. The toxic response is usually measured within 24 hours from the start of the test and can be conducted for a 4-day acute test or a 7- to 11-day chronic test. A measurable reduction or inhibition in bioluminescence indicates an adverse effect. The cost of the QwikSed analyzer (Sealite Instruments, Inc., Ft. Lauderdale, FL) and supporting software is approximately $15,000. The data from the QwikSed bioassay can be correlated with more conventional toxicity tests such as amphipods and sea-urchin development.

RAPID CHARACTERIZATION TOOLS IN THE ERA PROCESS

The Chief of Naval Operations (CNO) Policy for conducting ERAs identifies a three-tiered approach that incorporates different levels of assessment complexity.

- Tier 1 - Screening Risk Assessment (SRA) (Steps 1 and 2)
- Tier 2 - Baseline Ecological Risk Assessment (BERA) (Steps 3 to 7); and
- Tier 3 - Evaluation of Remedial Alternatives (Step 8)

This approach, which is consistent with the EPA Superfund Interim Final Ecological Risk Assessment Guidance for Superfund [6] consists of eight steps (Figure 1). RSC tools can be used to assist several step of this process.

Screening Risk Assessment

The goal of a Screening Risk Assessment (SRA) is to determine whether an exposure pathway is present between each chemical of interest and selected ecological receptors and to estimate risks for those chemicals for which pathways are identified. Such an assessment should employ existing data, and should not require additional data collection. Site data, however, do not always exist. If data are lacking, rapid characterization can map the extent of contamination in order to guide sampling for full contaminant of potential ecological concern (COPEC) analysis. By using

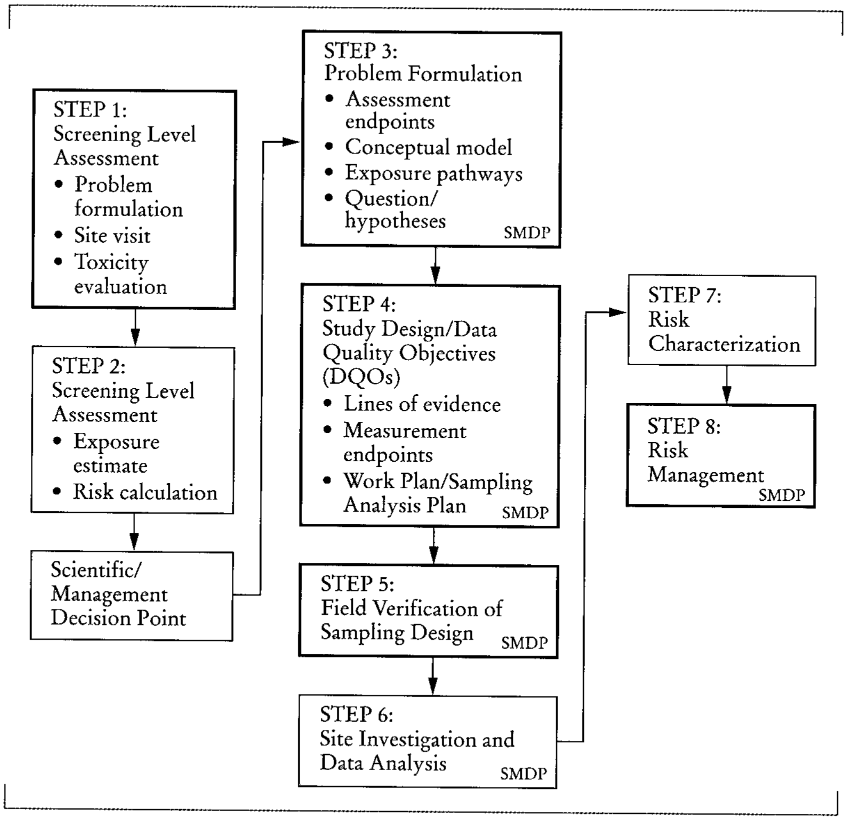


FIGURE 1. Navy Ecological Risk Assessment approach. Highlighted boxes indicate steps in which RSC tools can be used to facilitate the process.

RSC data to quickly map the area under investigation, subsequent sampling for full COPEC analysis can be more focused.

Baseline Ecological Risk Assessment

A Baseline Ecological Risk Assessment (BERA) is typically the most extensive activity within the ERA process, in terms of data collection and analysis, cost, and effort. There are several steps within the BERA in which rapid characterization tools can play a critical role, including Problem Formulation, Study Design/Data Quality Objectives (DQO) Process, and Verification of Field Sampling Design.

For example, two RSC tools were used for a sediment screening study at Hunters Point Shipyard to support a BERA sampling design (Steps 4 and 5). Surface sediment samples were collected in a grid-pattern from 94 locations in the five offshore areas of concern. Samples were screened for PCBs and heavy metals using the immunoassay technique and XRF spectrometry, respectively, at the SSC San Diego laboratory. The results were used to refine the sampling design for a more detailed study of sediment chemistry, toxicity, and bioaccumulation. In particular, screening results were used to ensure that the baseline assessment study sampling stations spanned the entire range of contaminant concentrations and, therefore, represented the full range of potential exposure. Ten percent of the screening samples were submitted to a standard analytical laboratory in order to obtain a quantitative analysis of all contaminants of concern, verify screening results, and provide additional surface sediment data supporting the assessment study.

Plots of PCB and copper (Cu) results are shown from one of the five offshore areas of concern (Figure 2). These results indicate two potential source areas for elevated PCBs in these offshore sediments, one on the northeast side and one on the west side of the embayment. Although the northeast area may be impacted by Navy operations, the source area to the west is at a creek mouth with potential non-Navy contributions of the target analytes from upstream locations onto Navy property. In the case of Cu, one potential source is indicated on the northeast side of the embayment, again potentially related to Navy operations. The screening

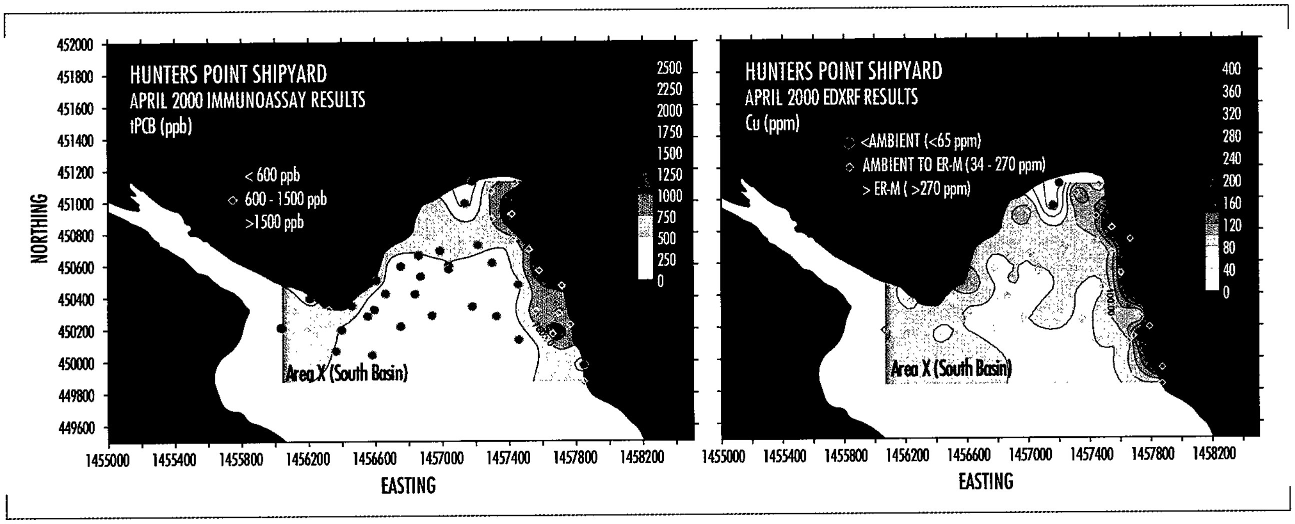


FIGURE 2. RSC tools implemented during BERA Steps 4 and 5 at Hunters Point Shipyard, CA. Immunoassay results for PCBs (left) and XRF results for Cu (right) are shown.

results can be used to delineate boundaries of impacted areas to ensure each potential source is sampled and laboratory data will be available to estimate relative source contributions to Navy sediments. As is often the case in sediment assessments, multiple potential sources are present. These sources need to be considered in the design of a sampling plan for the baseline assessment.

Evaluation of Remedial Alternatives

The purpose of this step is to ensure that remedial alternatives are adequately evaluated from an ecological perspective, so that the outcome of the remediation is not more detrimental to the environment than if the site had not been remediated [6]. Rapid characterization tools can play a role in this tier as well. If a remedial option is selected, costs are critically dependent on volumes or areas to be managed. Rapid characterization can be used to map out areas or volumes at higher density than were used for the assessment. Rapid characterization can also be used to verify the efficacy or completeness of a remedial option such as containment, cap or remove impacted sediments, and monitor the long-term efficacy and impact of management strategies.

CONCLUSIONS

A few important points must be considered in the selection and application of RSC tools to the ERA process. First, it is important that site-specific project goals and parameters as defined by the DQO Process must be considered. It is critical to ensure that the contaminants or criteria that are deemed to be decision drivers are detectable with the RSC tools that are available. Also, as with any method or technology, certain limitations exist. The primary limitations to RSC technologies are that they are often (1) non-specific, (2) semi-quantitative, and (3) matrix-sensitive. Because of these limitations, the data produced by RSC tools/ methods are not necessarily equivalent to those generated by standard methods. Depending on the data quality requirements established during the DQO Process, a well-designed RSC protocol, paired with laboratory validation, will be able to provide data that can be of sufficient quality and great value to the risk assessment. It is important to note that results can be misleading if non-equivalent data are combined without careful intercalibration. A few different approaches to the documentation and reporting of data can be used to avoid such problems when reporting results, particularly those from RSC methods. The first reporting approach is to always flag numbers generated by a non-standard method in spreadsheets and data reports, and to include text, references, or qualifiers that address any potential offsets from standard analyses. A second approach is to carry out site-specific calibration of RSC analyses and to report only corrected, calibrated data. A third option, particularly for RSC analyses that generate only qualitative data (i.e., data that identify the presence or absence of target analytes, but may have no relationship to true concentrations of the analytes) is not to report numerical values, but instead report qualitative values (e.g., non-detect, etc). Samples are either ranked or ranges are reported. Finally, a concern voiced by many potential users of RSC tools is that, since they are not subject to the same quality assurance/quality control (QA/QC) protocols and rigors as are standard procedures, they will make the user vulnerable by not standing up to regulatory or legal scrutiny. While these concerns are not trivial, it is clear that there are a growing number of case studies in which remedial project managers, regulators, and the user community have accepted RSC data as a critical, though not stand-alone, part of the analytical and decision-making process. In any case, the intent to use RSC tools, and how the resulting data will be interpreted and managed, should be addressed up front with regulators and other stakeholders.

Implementation of rapid characterization tools in ecological risk assessments will improve sampling and reduce uncertainty at several steps of the remedial investigation/feasibility study process without the enormous cost of traditional resampling efforts. Use of these tools moves the ERA process forward in the most time- and cost-effective manner with minimum uncertainty.

REFERENCES

1. Chief of Naval Operations (CNO) Letter 5090 Ser N453E/9U595355 dated 05 April 1999; Navy Policy for Conducting Ecological Risk Assessments.
2. United States Environmental Protection Agency. 1997. "Field Analytical Site Characterization Technologies, Summary of Applications," EPA-542-R-97-011.
3. California Military Environmental Coordination Committee. 1996. "Field Analytical Measurement Technologies, Applications and Selection," April.
4. United States Environmental Protection Agency. 2001. *Field Analytical Technologies Encyclopedia*, January, http://fate.clu-in.org/
5. Gilbert, R. O. 1987. *Statistical Methods for Environmental Pollution Monitoring*, Van Nostrand Reinhold, New York, chapter 9.
6. United States Environmental Protection Agency. 1997. "Ecological Risk Assessment Guidance for Superfund: Process for Designing and Conducting Ecological Risk Assessments—Interim Final," EPA 540-R-97-006.
7. United States Environmental Protection Agency. 1998. "Method 6200: Field Portable X-ray Fluorescence Spectrometry for the Determination of Elemental Concentrations in Soils and Sediments," Revision 0, January.
8. Filkins, J. 1992. "Draft Experimental Methods: Estimates of PAHs in Lacustrine Sediment by Fluorometry," USEPA—Large Lakes Research Station, unpublished interim report.
9. United States Environmental Protection Agency. 1996. "Region I, EPA—New England: Immunoassay Guidelines for Planning Environmental Projects," October.
10. Giesy, J., C. Rosiu, R. Graney, and M. Henry. 1990. "Benthic Invertebrate Bioassays with Toxic Sediment and Pore Water," *Environmental Toxicology and Chemistry*, vol. 9, pp. 233–248.
11. American Society for Testing and Materials (ASTM). 1999. "Standard Guide for Conducting Toxicity Tests with Bioluminescent Dinoflagellates," 1999 *Annual Book of ASTM Standards*, vol. 11.05, pp.1467–1477.

❖

Victoria J. Kirtay

MS in Environmental Science and Technology, Bosporus University, Istanbul, Turkey, 1994

Current Research: Contaminated marine sediment management; field-screening sensors.

Sabine E. Apitz

Ph.D. in Oceanography, Scripps Institution of Oceanography, University of California at San Diego, 1991

Current Research: Contaminant–sediment interactions, environmental assessment, marine sediment management technology and policy.

LIST OF TRADEMARKS

Alphatech® is a registered trademark of Alphatech, Inc.

Bell Atlantic® is a registered trademark of Bell Atlantic Corporation.

Benthos® is a registered trademark of the Benthos, Inc.

Cyc® is a registered trademark of Cycorp, Incorporated.

eBay™ is a trademark of eBay Inc.

Freewave® is a registered trademark of FreeWave Technologies, Inc.

Hamamatsu® is a registered trademark of the Kabushiki Kaisha Corporation.

HotJava® is a registered trademark of Sun Microsystems, Inc.

IBM® is a registered trademark of the IBM Corporation.

Information Science Institute® is a registered trademark of the Institute for Information Sciences, Inc.

Java™, Java2™, J2EE™, and Solaris™ are trademarks of Sun Microsystems, Inc.

JavaScript® is a registered trademark of Sun Microsystems, Inc.

Jeronimo® is a registered trademark of Appian Graphics Corporation.

LabView® is a registered trademark of National Instruments Corporation.

Microtox® is a registered trademark of AZUR Environmental Corporation.

NetMeeting®, PowerPoint®, Windows®, and Windows NT® are registered trademarks of the Microsoft Corporation.

Netscape™ is a trademark of Netscape Communications Corporation.

Pentium® is a registered trademark of the Intel Corporation.

Quava™ is a trademark of Science Applications International Corporation.

Rational Rose® is a registered trademark of Rational Software Corporation.

SPARC® is a registered trademark of SPARC International Inc.™
Products bearing SPARC® trademarks are based on an architecture developed by Sun Microsystems, Inc.

SRI International® is a registered trademark of SRI International.

Teknowledge® is a registered service mark of Teknowledge, Inc.

Texas Instruments® is a registered trademark of Texas Instruments Incorporated.

AFTERWORD

AUTHOR INDEX